REMARKS

I. Introduction

Applicant respectfully requests reconsideration of the present application in view of the foregoing amendments and in view of the reasons that follow.

Claims 12-15, 22-25, 27-28, 30-31, 33-34, 36-37, 44-47, 54-57, 64-67 and 70-92 are canceled. The cancellation of claims does not constitute acquiescence in the propriety of any rejection set forth by the Examiner. Applicant(s) reserve the right to pursue the subject matter of the canceled claims in subsequent divisional applications.

Claims 1-5 and 68 are currently amended. Claims 6-11, 16-21, 26, 29, 32, 35, 38-43, 48-53, 58-63 and were withdrawn.

This amendment adds, changes and/or deletes claims in this application. A detailed listing of all claims that are, or were, in the application, irrespective of whether the claim(s) remain under examination in the application, is presented, with an appropriate defined status identifier.

Upon entry of this Amendment, claims 1-21, 26, 29, 32, 35, 38-43, 48-53, 58-63, 68-69 and 93-95 will remain pending in the application.

Because the foregoing amendments do not introduce new matter, entry thereof by the Examiner is respectfully requested.

II. Response to Issues Raised by Examiner in Outstanding Office Action

a. Objection to Oath/Declaration

The Office maintains an objection to the oath based on non-initialed or non-dated alterations. To support this contention the Office cites 37 CFR 1.52(c). This rule provides:

> (c)(1) Any interlineation, erasure, cancellation or other alteration of the application papers filed must be made before the signing of any accompanying oath or declaration pursuant to § 1.63 referring to those application papers and should be dated and initialed or signed by the applicant on the same sheet of paper. Application papers containing alterations made after the signing of an oath or declaration referring to those application papers must be supported by a supplemental oath or declaration

> under § 1.67. In either situation, a substitute specification (§ 1.125) is required if the application papers do not comply with paragraphs (a) and (b) of this section.
> (2) After the signing of the oath or declaration referring to the application papers, amendments may only be made in the manner provided by § 1.121.
> (3) Notwithstanding the provisions of this paragraph, if an oath or declaration is a copy of the oath or declaration from a prior application, the application for which such copy is submitted may contain alterations that do not introduce matter that would have been new matter in the prior application.

Applicants note that rule 1.52 requires pages with changes to be dated and signed by the Applicant. The changes to the declaration occur in the signature block and are immediately signed and dated below the changes. As the inventor has signed and dated the page with the changes, the authenticity of the changes is not in doubt and Applicants request withdrawal of this objection.

b. Claim Rejections - 35 U.S.C. § 112, First Paragraph

Claims 68-69 and 93 are rejected under 35 U.S.C. § 112, second paragraph, as containing subject matter which was not described in the specification in such a way as to reasonably convey to one skilled in the relevant art that the inventor(s), at the time the application was filed, had possession of the claimed invention. The Office maintains, "The claimed invention is directed to an isolated or purified peptide (SEQ ID NO:3), that has at least 80% or at least 90% sequence identity to the peptide as claimed. The claims encompass a genus of variants that are highly variable. A skilled artisan cannot envision the detailed chemical structures for all the variants encompassed by the claims. Office Action, p. 4.

Applicants disagree with the Office. However, in order to further prosecution, Applicants have canceled claims 69 and 93 providing for percentage identity with the peptides of claim 1. Claim 68 is drawn to peptides with 1-5 conservative amino acid substitutions. Support for this claim is found in paragraph [0029] and original claim 68. Additionally, paragraph [0056] describes conserved amino acid substitutions involve replacing one or more amino acids of the protein of the invention with amino acids of similar size, size and/or hydrophobicity characteristics. As of the time of filing, groups of amino acids considered appropriate for substitutions had been well studied and understood by those of skill in the art. For example, William Taylor, "The Classification of Amino Acid Conservation", *J. Theor. Biol.* 119,205-218 (1986), and Bordo, et al, *J. Mol. Biol.*, 217,721-

729 (1991) (See Attached), both describe the classification of amino acids well before the filing date of the application. A person of skill in the art could readily identify conservative amino acids from these applications and common knowledge at the time of filing. Based on these disclosures, replacement of 1-5 amino acids with conservative substitutions is described in the application and a person of skill in the art would recognize that Applicants were in possession of the claimed invention at the time of filing.

Claims 68-69 and 93 are rejected under 35 U.S.C. § 112, first paragraph, because the specification, while being enabling for the proteins set for in SEQ ID NO: 3, does not reasonably provide enablement for any peptide having at least 80% or 90% sequence homology to SEQ ID NO: 3.

As noted above, Applicants disagree with the Office. However, in order to further prosecution, Applicants have canceled claims 69 and 93 providing for percentage identity with the peptides of claim 1. Claim 68 is drawn to peptides with 1-5 conservative amino acid substitutions.

The fact that experimentation may be complex does not necessarily make it undue, if the art typically engages in such experimentation. *In re Certain Limited-Charge Cell Culture Microcarriers*, 221 USPQ 1165, 1174 (Int'l Trade Comm'n 1983), *aff'd. sub nom., Massachusetts Institute of Technology v. A.B. Fortia*, 774 F.2d 1104, 227 USPQ 428 (Fed. Cir. 1985). See also *In re Wands*, 858 F.2d at 737, 8 USPQ2d at 1404.

As long as the specification discloses at least one method for making and using the claimed invention that bears a reasonable correlation to the entire scope of the claim, then the enablement requirement of 35 U.S.C. 112 is satisfied. *In re Fisher*, 427 F.2d 833, 839, 166 USPQ 18, 24 (CCPA 1970). A person of skill in the art would be able to practice the claimed invention using methods described within the specification.

Although the Office maintains that substitution of conserved amino acids is not enabled by the specification, it is unclear what aspect of the experimentation is "undue experimentation" under the current case law for the standard of enablement. The currently claimed invention is to short isolated peptides and methods of utilizing these peptides.

Preparation of such peptides is well known to one of skill in the art in a multitude of techniques including standard peptide synthetic technology. The Specification outlines all of the methods and tests necessary in order to use these peptides for testing activity. See the methods outlined in the Application and used for some of the described peptides in Examples 1-14. The specification provides sufficient guidance for following the claimed methods to determine if a peptide affects the rate of degradation of type II collagen or the rate of chondrocyte hypertrophy. As noted above, as long as the specification discloses at least one method for making and using the claimed invention that bears a reasonable correlation to the entire scope of the claim, then the enablement requirement of 35 U.S.C. 112 is satisfied. *In re Fisher*, 427 F.2d 833, 839, 166 USPQ 18, 24 (CCPA 1970). Applicant respectfully requests reconsideration and withdrawal of the rejection.

c. Claim Rejections - 35 U.S.C. § 102

Claims 1, 69, and 93 are rejected under 35 U.S.C. § 102(b) as being anticipated by Qvist et al. (US Patent No. 6,110,689, August 29, 2000) and Claims 1, 68-69, and 93 over Shriners Hospitals for Crippled Children (WO 94/14070, 1994), cited on IDS filed April 30, 2004, based on the open language in the claim which reads on all of SEQ ID NO: 3 embedded in a longer sequence. Office Action, pp. 11-12 and 13-14. Applicants have canceled claims 69 and 93. Regarding claim 1, Applicants have amended the claims to recite isolated peptides consisting of the provided amino acid sequences. The claims do not encompass other peptides comprising SEQ ID NO: 3. If Applicant's explanation and amendments are not sufficient to satisfy the Office, Applicant requests clarification regarding an Amendment to specify this understanding.

In addition, there is no disclosure, teaching or suggestion in Qvist of isolated or purified peptides of SEQ ID No:3, the hydroxylated versions of this peptide with an additional glycine at the N- terminus (SEQ ID Nos: 6-9), or any of the other peptides of the present invention.

Likewise, WO 94/14070 does not teach or suggest an "isolated or purified" peptide of SEQ ID No: 3, nor the other peptides of in claim 1. Moreover, there is no teaching or suggestion in either of these references for a person of ordinary skill in the art to conclude

that the peptides of the present invention would modulate and regulate cell differentiation as well as the degradation of collagen.

Applicant respectfully requests reconsideration and withdrawal of the rejection.

CONCLUSION

The present application is now in condition for allowance. Favorable reconsideration of the application as amended is respectfully requested.

It is acknowledged that the foregoing amendments are submitted after final rejection. However, because the amendments do not introduce new matter or raise new issues, and because the amendments either place the application in condition for allowance or at least in better condition for appeal, entry thereof by the Examiner is respectfully requested.

The Examiner is invited to contact the undersigned by telephone if it is felt that a telephone interview would advance the prosecution of the present application.

The Commissioner is hereby authorized to charge any additional fees which may be required regarding this application under 37 C.F.R. §§ 1.16-1.17, or credit any overpayment, to Deposit Account No. 19-0741. Should no proper payment be enclosed herewith, as by a check being in the wrong amount, unsigned, post-dated, otherwise improper or informal or even entirely missing, the Commissioner is authorized to charge the unpaid amount to Deposit Account No. 19-0741. If any extensions of time are needed for timely acceptance of papers submitted herewith, Applicant(s) hereby petition(s) for such extension under 37 C.F.R. §1.136 and authorizes payment of any such extensions fees to Deposit Account No. 19-0741.

Respectfully submitted,

Date Feb. 12, 2007

By _____

FOLEY & LARDNER LLP
Customer Number:
22428
PATENT TRADEMARK OFFICE
Telephone: (202) 672-5569
Facsimile: (202) 672-5399

Stephen B. Maebius
Attorney for Applicant
Registration No. 35,264

J. theor. Biol. (1986) **119**, 205–218

The Classification of Amino Acid Conservation

WILLIAM RAMSAY TAYLOR

Laboratory of Molecular Biology, Dept of Crystallography, Birkbeck College, London WC1E 7HX, U.K.

(*Received* 20 *September* 1985, *and in revised form* 1 *November* 1985)

A classification of amino acid type is described which is based on a synthesis of physico-chemical and mutation data. This is organised in the form of a Venn diagram from which sub-sets are derived that include groups of amino acids likely to be conserved for similar structural reasons. These sets are used to describe conservation in aligned sequences by allocating to each position the smallest set that contains all the residue types brought together by alignment. This minimal set assignment provides a simple way of reducing the information contained in a sequence alignment to a form which can be analysed by computer yet remains readable.

1. Introduction

The high level of expertise current in nucleic acid research has led to the revelation of large numbers of protein sequences and the ability specifically to alter these in a controlled way. New sequences often exhibit a close homology with proteins which have had their structure determined crystallographically and using advanced computer graphic facilities it is possible theoretically to alter the amino acid side chains of the known structure to represent the sequence of unknown structure (e.g. Blundell *et al.*, 1983). The new techniques are used increasingly to design proteins with altered properties using the methodology of site-specific mutagenesis.

These activities require a good understanding of the basic principles of protein structure and, in particular, it is necessary to anticipate the structural effect of introducing a new amino acid into a known structure. This assessment is often based on the likelihood matrix of amino acid mutabilities derived by Dayhoff *et al.* (1972, 1978) or on the number of nucleotide base changes required to effect the substitution (Fitch, 1966). Such measures, however, ignore aspects of the substitution that are relevant to the local structural environment or known function of the residue and in building hypothetical structures and designing new mutants it is these details which are important.

In this paper I consider measures of amino acid relatedness in common use with the aim of extracting from them features which will best assist a protein engineer faced with the problem of making a mutation and assessing a sequence alignment. These features are represented as groupings of amino acids (sets) which is a form that retains a descriptive quality yet allows quantitative manipulations using the formalism of set logic.

0022-5193/86/060205+14 $03.00/0

2. Measures of Amino Acid Relatedness

(A) MUTATION DATA

For every pair of the 20 naturally occurring amino acids Dayhoff *et al.* (1972) have determined the probability (or odds) that the mutation will occur in either direction. This matrix was most clearly presented by Sander & Schulz (1979) (see also Schulz & Schirmer, 1979) in a form where the entries have been ordered to bring frequently exchanging amino acids together. Even in its ordered form it is still difficult fully to appreciate the information contained in Dayhoff's matrix. However, using the technique of multi-dimensional scaling, French & Robson (1983) reduced the matrix to a two dimensional plot, in which frequently exchanging amino acids are closest together (see Fig. 1(a)). A similar diagram (Fig. 1(b)) was produced by minimising the deviation from a 2-D structure in which the entries of Dayhoff's matrix represent the inverse of ideal target distances between pairs of amino acids (Taylor, 1981). Both these figures are roughly elliptical and projecting the amino acids onto the circumference of each ellipse produces an even simpler representation with no great loss of information. In this form the long axis of the ellipse corresponds to molecular volume while the short axis corresponds to hydrophobicity. These simplifications are presented in Figs 1(c) and 1(d). The cyclic order of amino acids obtained from this simplification is almost the same as that derived by Swanson (1984) (see Fig. 1(e)).

All the above representations of Dayhoff's matrix indicate that it can be largely accounted for by the effect of only two determining factors; hydrophobicity and size. This remarkable observation must obviously dominate any attempt to codify amino acid conservation.

(B) PHYSICAL DATA

Physico-chemical properties

The wide variety of physico-chemical properties manifest in the amino acid side-chains has been thoroughly considered by Sneath (1966). These have been adopted, or deduced independently, by others and used as a basis for considering relatedness between protein sequences. McLachlan (1972) summarizes these relationships by assigning a value to each transition between pairs of amino acids. These scores are presented graphically in Fig. 2 using the circle of amino acids derived from Dayhoff's matrix as a frame on which lines connect related amino acids. All high scoring transitions and most other connections on the graph are local, indicating agreement between chemistry and mutability. The less local connections mainly join hydrophobic residues, which, with the exception of proline, are relatively adjacent on the less idealised representations of Dayhoff's matrix (Fig. 1). The missing links are, perhaps, more revealing: there is only a weak link between Tyr and Trp which are strongly tied in Dayhoff's mutation matrix and there is no connection between the adjacent negatively charged residues and Gly and Pro.

Measurement of properties cannot determine a common scale without reference to protein sequences and structures. The idea of such a scaling can be appreciated

Fig
dimen
excha
point
distan
with a
ideali
recon
Swan
from
characteri
Gly G, A
Phe F, Ty

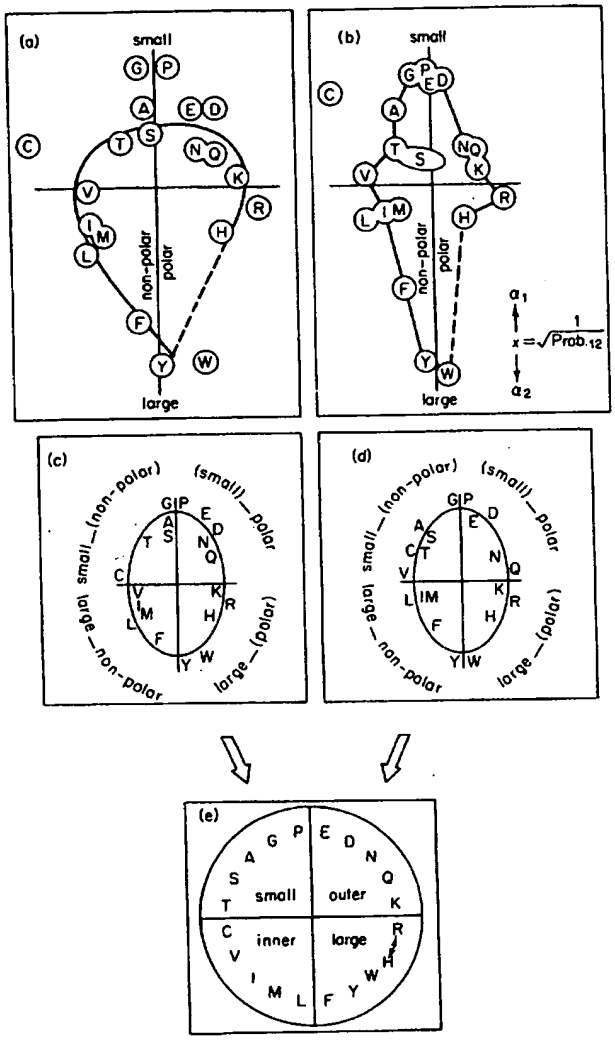


FIG. 1. Representations of Dayhoff's mutation odds matrix. (a) Projection of the matrix by multi-dimensional scaling (adapted from Robson & French (1983)). Amino acids which are close together exchange frequently. (b) The equivalent diagram to (a) produced by pseudo-energy minimization to a point at which each amino acid lies in a position which gives the minimum sum of squares over the distance equation shown. (c) and (d) Idealizations of (a) and (b) respectively. The properties associated with each quadrant are indicated with the property of lesser importance bracketed. (e) A further idealization of the two plots which are constrained to a circle. Ambiguities in the cyclic order have been reconciled by consideration of both original plots. The resulting order agrees with that obtained by Swanson (1984) except for the exchange of Arg and His as indicated by an arrow. (This probably arises from Swanson's use of the Dayhoff (1978) revised matrix). Swanson's nomenclature for the quadrant characteristics is also indicated. The one letter and three letter codes for the amino acids are as follows: Gly G, Ala A, Val V, Leu L, Ile I, Ser S, Thr T, Asp D, Glu E, Asn N, Gln Q, Lys K, His H, Arg R, Phe F, Tyr Y, Trp W, Cys C, Met M, Pro P, asx b, glx z.

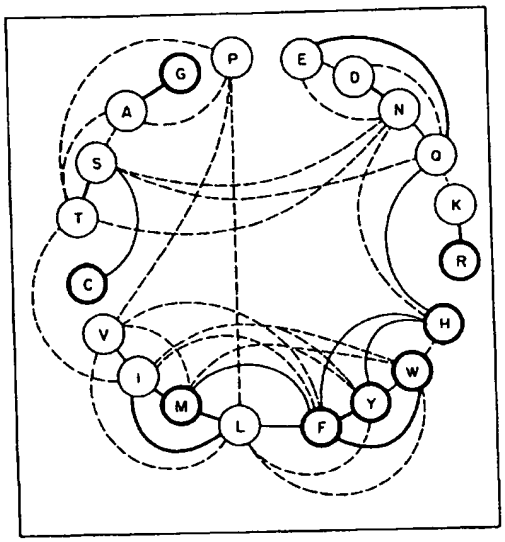


FIG. 2. Chemical relatedness of the amino acids as quantified by McLachlan (1972) displayed on the circle of residues derived from Dayhoff's matrix (Fig. 1(e)). The degrees of expected conservation are indicated as follows (with McLachlan's score in brackets). Strong conservation of type = heavy circle (6); Moderate conservation of type = lighter circle (5); Strong conservation of properties = heavy line (3); Moderate conservation of properties = finer line (2); Weak conservation of properties = broken line (1). Most connections are local (i.e. chemistry agrees with mutation data) with the exception of long links from hydrophobic residues to Pro and Ser & Thr to Asn & Gln and the absence of links between Gly & Pro and Glu & Asp.

from the relative lengths of the ellipses derived from Dayhoff's matrix (see Figs. 1(c) and 1(d)) in which the longer axis is associated with change of size indicating that this, on average, is dominant over hydrophobicity. Grantham (1974) scaled three properties to McLachlan's (1971) matrix of amino acid substitution frequencies. These were volume, polarity and the fraction of carbon in the side-chain (composition). He found the most dominant property was again volume followed by polarity.

Secondary structure propensities

From statistical analysis of the sequences of proteins of known structure, propensities to adopt a secondary structure have been determined for each amino acid (e.g. Chou & Fasman, 1974; Garnier *et al.*, 1979). These preferences have been used to account for clusters of amino acids which are unexpected on a physico-chemical basis. A clear example is the close association of G, P, D and E (see Fig. 1). These associate because of a propensity to lie in sharply turning regions on the surface of the protein. Gly, because of the flexibility it imparts to the local chain; Pro because of the built in turn configuration created by its back-bonding side chain; and Asp and Glu because of a requirement to expose their charges to solvent. Robson & French (1983) indicate other instances including the close association of

Glu and Ala both of which favour α-helical structure, and the tight association of L, I, M, V and to a lesser extent, F and Y which tend towards β-structure.

3. Venn Diagram of Amino Acid Sets

The idea of using a Venn diagram to represent the different relationships among the amino acids was adopted from Dickerson & Geis (1969) and extended to incorporate some of the observations discussed above. The overall layout of the diagram (Fig. 3(a)) was based on the 2-D arrangement derived from Dayhoff's mutation matrix. Amino acids were then displaced (by as little as possible) from this arrangement to form groups of residues related by common physio-chemical properties.

(A) SIZE AND HYDROPHOBICITY

The major sets group the amino acids by size and hydrophobicity: both properties which were seen to dominate the structure of Dayhoff's Matrix. Two overlapping sets were used to describe hydrophobicity. One was defined as all amino acids which have a polar group in their side-chain and is referred to as *polar*. The second group is less well defined and contains the amino acids which were considered to be hydrophobic. This set contains some amino acids which have polar side-chains. These consequently lie in the intersecting region of the two sets which can be considered the set of amino acids which are ambivalent to water. The inclusion of Lys in this set is justified by its long aliphatic side-chain which has been observed (Cohen *et al.*, 1982) to extend from a buried location and expose the terminal charge to solvent.

The location of Pro in Fig. 1 conflicts with its hydrophobic character. It was, thus, left unclassified by the two sets *hydrophobic* and *polar*. Similarly, Gly is often considered to lack a side-chain and be consequently unclassifiable in hydrophobic terms (e.g. Rees & Sternberg 1984). However, as Gly is often found buried in the interior of proteins it was classified as hydrophobic.

The volume of the side-chain was considered to be sufficiently important to justify classification by two sets. The larger of these, called *small*, contains the nine smallest amino acids by side-chain volume (Klapper, 1971) each less than 60 Å^3. A subset of this, called *tiny*, includes the four acids with less than three (non-H) side-chain atoms all of which are smaller than 35 Å^3.

The relationship of Cys to the sets defined above is rather ambiguous. Although its side-chain has only two atoms, the sulphur atom is relatively large, placing it on the *Tiny–Small* borderline. Its classification is further complicated by the occurrence of the sulphur in two oxidation states. The reduced form contains a polarizable S–H bond which suggests a similarity to Ser (O–H) and with which it is associated in McLachlan's table (Fig. 2). On formation of a disulphide bond, however, this property is lost, placing the residue more firmly in the hydrophobic camp. The associated loss of conformational freedom is difficult to assess but may be associated with an effective increase in volume as the linked residue cannot accommodate as easily to structural fluctuations. Poor packing in the hydrophobic core of the

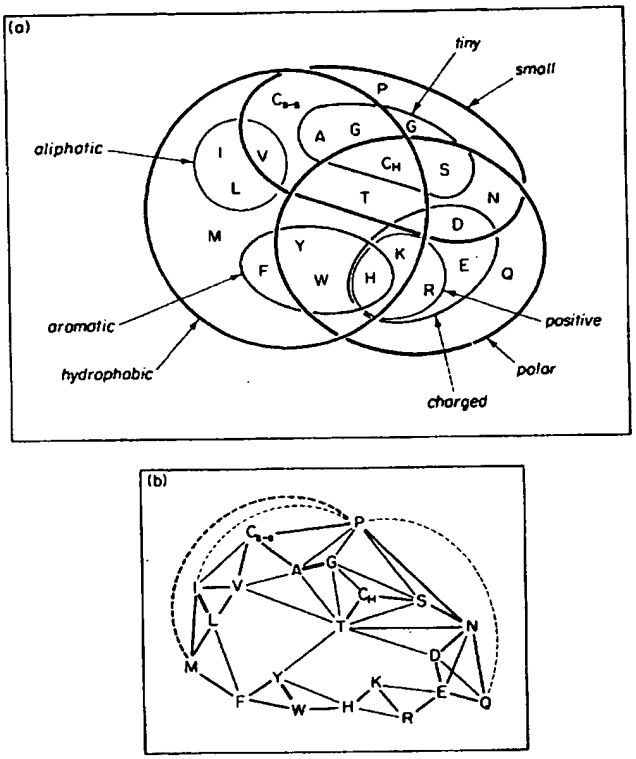


FIG. 3. (a) The Venn diagram shows the relationship of the 20 naturally occurring amino acids to a selection of physio-chemical properties which are important in the determination of protein tertiary structure. The diagram is dominated by properties relating to size and hydrophobicity. The amino acids are divided into two major sets, one containing all amino acids which contain a polar group (*polar*) and a set which exhibit a hydrophobic effect (*hydrophobic*). A third major set, *small*, is defined by size and contains the nine smallest amino acids. Within this is an inner set of smaller residues, *tiny*, which have at most two side-chain atoms. The location of Cys is ambiguous as the reduced form (CH) has similar properties to serine, while the oxidized form (Css) may be more equivalent to Val. Other sets include full-charge (referred to as *charged*) which contains the subset *positive* (negative is defined by implication) and *aromatic* and *aliphatic*. The latter set is not as general as the name implies and contains only residues containing a branched aliphatic side-chain. Because of its unique backbone properties, proline was excluded from the main body of the diagram. An equivalent exclusive position is suggested for Gly by a small G. (b) An alternate representation of the relationships. A network of amino acids is formed by connecting those which differ by no more than two properties in the Venn diagram. Pairs which share the same subset are connected by a heavy bar, those with only one different property by a bold line and those which differ by two properties by a fine line. To improve clarity a few of the latter connections are omitted. Many of these only connect to one of an unresolved pair (e.g. I and L), and it can be assumed that the connection is made to both. Because of its unique position, Pro is able to make quite long range connections some of which are indicated by broken lines.

immunoglobulins near the intra-molecular disulphide link is suggestive of this (Cohen *et al.*, 1981, Taylor, 1981). Considering all these aspects, two locations are suggested for Cys on the Venn diagram. However, a location in the sub-set containing Thr is also possible.

(B) OTHER SETS

The remaining set allocations are based on obvious physico-chemical properties. These include *aromatic* (ring containing side-chains) and *aliphatic*. The latter set, however, includes only amino acids with branched aliphatic side-chains and largely reflects the frequency with which this type of residue is found in β-pleated sheet structure.

The set of *charged* amino acids contains only those which are normally (or often) fully ionized. The subset *positive* is included, with *negative* defined by implication.

For simplicity, as few sets as possible were introduced, yet even these produce almost complete segregation of the amino acids with only Y-W, I-L and A-G grouped in the same sub-sets. Additional sets can easily be imagined but these often only create further distinction between residues which are already segregated. However, an important property not well represented is hydrogen-bonding ability. To distinguish the sets of hydrogen-bond donors and acceptors on the Venn diagram (Fig. 4) would greatly reduce clarity they are thus indicated separately in Figs 4(c) and 4(d).

(C) NETWORK REPRESENTATION

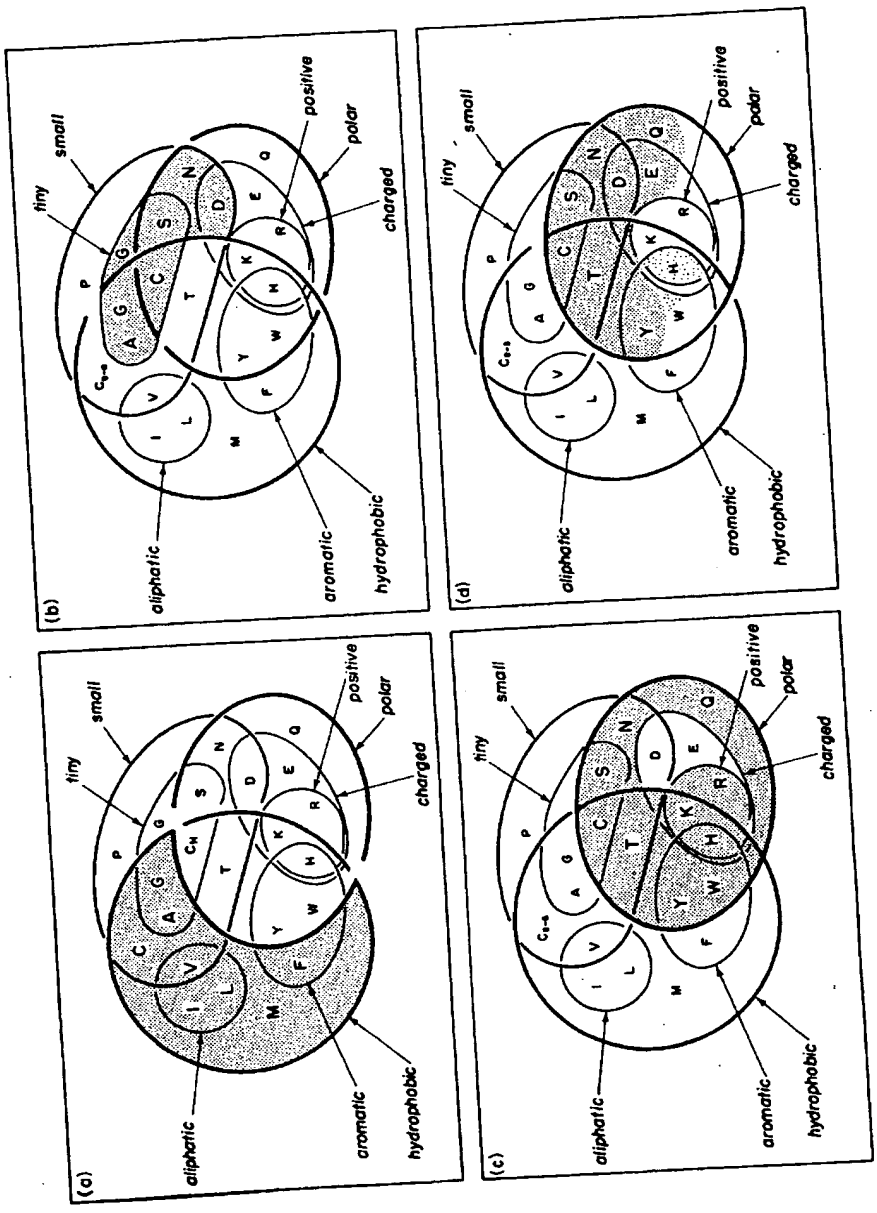
A useful representation of the Venn diagram can be made by removing the set boundaries and connecting adjacent residues to form a network (Fig. 3(b)) in which no connected pair differ by more than two properties. In this form the structure is more easily compared to other representations and the circular arrangement of amino acids corresponding to Fig. 1 is readily apparent.

The most significant deviation from the form of Dayhoff's matrix is the separation between the negatively charged amino acids and Pro and Gly. This feature corresponds more to the physico-chemical relationships defined by McLachlan (Fig. 2). However, leaving Pro (and perhaps Gly) unclassified with respect to hydrophobicity allows connections to be made with both hydrophilic (S, N, Q) and hydrophobic residues (F, M, L, I, V and A). Some of these longer connections are indicated in Fig. 3(b) by broken lines.

It is also possible to formally reduce the Venn diagram (or network) to a tree structure of the type described by Jiménez-Montaño (1984). However, unless the corresponding tree contained multiple amino acid entries information in the Venn diagram would be lost.

4. Minimal Set Assignment

The Venn diagram (Fig. 3(a)) represents a compromise between mutation data and chemistry. A similar compromise might have been achieved in a less graphical way by calculating a distance matrix of the type described by Grantham (1974). Such tables of relatedness are useful for considering amino acid changes without regard for the local environment in the protein structure in which the change occurs (e.g. exposure to solvent, secondary structure, etc.) and, consequently, can be applied uniformly along any pair of sequences to produce an overall measure of relatedness



between
ingly co
This requ
environm
to tackle
be previ
would b
approach
With
many re
can, of
and com
obscures
intereste
time for
With
the Ven
find the
together
Applyin
a qualit
of amin
conserv
twenty
Desp
is very
it is dif
the co
H, K,
which
of the
examp
except

FIG.
(Fig. 3
hydrop
nomen
assigne
sub-se
polar
the m
hydrog
It sho
receiv
state a
uncert
importanc

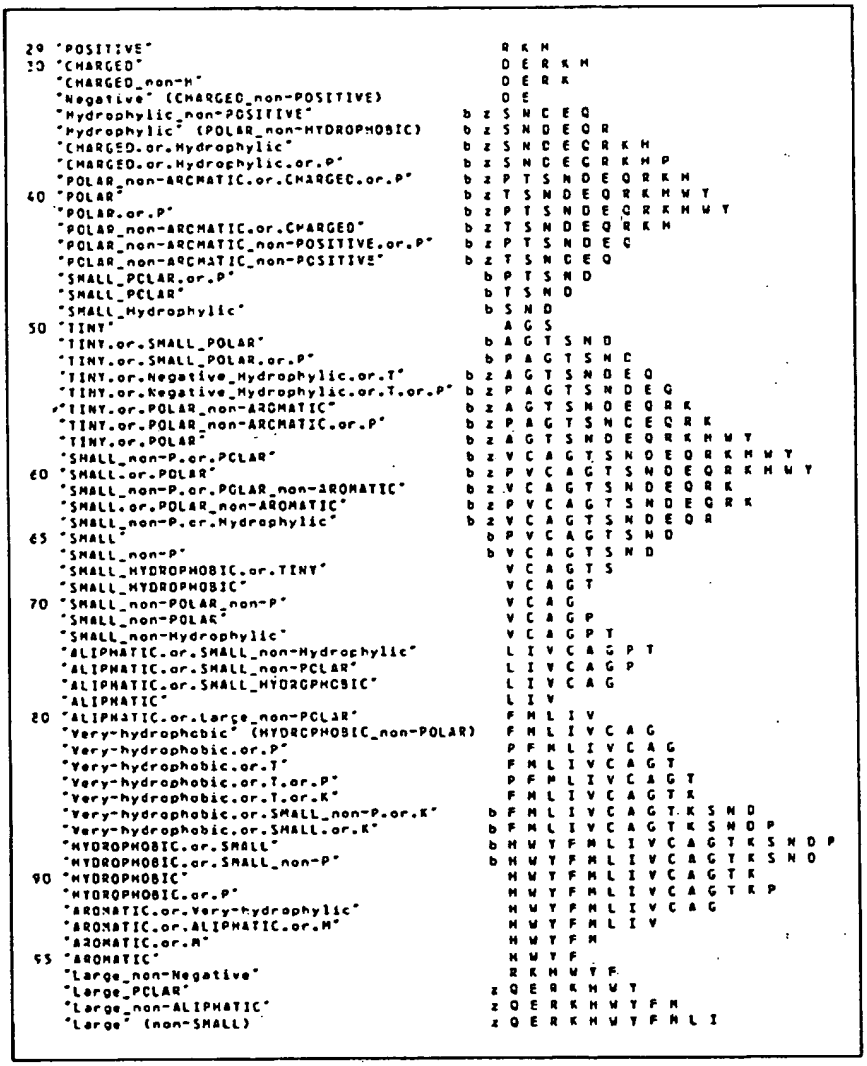
between them. With greater knowledge of protein structures, however, it is increasingly common that substitutions are considered in a sequence of known structure. This requires a new approach to assessing the substitution odds for a given structural environment. The use of tables of relatedness, moreover, cannot be easily adapted to tackle this problem as every environment would require its own table which must be previously calculated from substitutions occurring in that environment. This would be a very useful set of tables to have but for present work a more flexible approach is required.

With the vast increase in known protein sequences it is now common to have many related sequences to compare and not simply a pair. This type of problem can, of course, be tackled using a distance measure for all or part of a sequence and compiling a table of homology for each pair of sequences. The result, however, obscures the qualitative description of the conservation and workers currently interested in this type of problem are often nucleic-acid biochemists who have little time for numerical abstractions.

With these problems in mind, a method was devised to use the sets defined by the Venn diagram (Fig. 3(b)) to describe sequence alignments. This was, simply, to find the smallest set or sub-set which includes all the amino acid types brought together by alignment. The resulting set will be referred to as the minimal set. Applying this operation over the whole length of the aligned sequences produces a qualitative description of the conservation at every point. Furthermore, the number of amino acids which constitute the assigned set give a measure of the degree of conservation at that point ranging from one (for absolute conservation of type) to twenty (the set of all amino acids).

Despite the few sets included in the Venn diagram, the number of possible subsets is very large and includes many which have little physical meaning. For example; it is difficult to imagine a structural environment or function which would promote the conservation of the set formed by the union of *aliphatic* and *positive* (V, I, L, H, K, R). A list of almost seventy sets and subsets has, thus, been compiled which might be maintained by structural selection pressures. These naturally consist of the union and intersection of sets which overlap in the Venn diagram (graphical examples of two of these are shown in Figs 4(a) and (b)). One set which is an exception to this was created to reflect the conservation of Gly, Pro and the negatively

FIG. 4. (a) and (b) indicate how two subsets are derived from the sets indicated on the Venn diagram (Fig. 3(a)). (a) is formed by the amino acids that are *hydrophobic* but not *polar* (or in set nomenclature: *hydrophobic* $\wedge \sim$ *polar*, where $\sim$ indicates negation and $\wedge$ indicates set intersection). In the alternative nomenclature defined in the legend to Fig. 5 this becomes *hydrophobic* non-*polar* but as it is a commonly assigned set it is given the "trivial" name of *very-hydrophobic*. (b) illustrates a more complex derivative sub-set formed by *tiny* $\cup$ (*small* $\wedge$ (*polar* $\wedge \sim$ *hydrophobic*)), where $\cup$ indicates set union. In Fig. 5 *polar* $\wedge \sim$ *hydrophobic* is given the trivial name of *hydrophylic* and union is indicated by . or., producing the more comprehensible name of *tiny*. or. *small*_*hydrophylic*. (c) and (d) indicates the two sets of hydrogen-bond donors (c) and acceptors (d). Their definitions were taken from Baker & Hubbard (1984). It should be noted, however, that hydrogen bonds involving Cys are rare and that Met might possibly receive a hydrogen-bond. Also, the classification of His is complicated by its frequent change of ionisation state as only the unprotonated state can receive a hydrogen bond. For clarity, and as there is still some uncertainty in these sets, they were not included in the main Venn diagram despite their obvious structural importance.



charged amino acids in the bend regions of protein structures. Other minor sets consisting of closely related pairs, including Ser and Thr, Phe and Tyr, and Arg and Lys, were also added. All these sets are defined in Fig. 5 where they are described using a nomenclature adapted from set logic.

A few example applications of the minimal set assignments to aligned sequence fragments are shown in Fig. 6. To give some impression of how set assignment varies with the number of aligned sequences and the overall homology of the sequences, a progression is shown from a few closely related sequences of a particular immunoglobulin domain to an extended alignment of the same domain (Fig. 6). In these it is clear that conservation is maintained mainly in the regions of secondary structure. This observation can be quantified by plotting the set size at each position in the alignment. In Fig. 7 this is done for different numbers of closely related sequences of the immunoglobulin κ-chain light-variable domain. Of these sequences the Bence-Jones protein REI has a known crystallographic structure (Epp *et al.*, 1975). The strands of β-structure found in this protein lie in two sheets which stack together like a sandwich (Cohen *et al.*, 1981). One side of each sheet is buried while the other is exposed to solvent. As the amino acid side-chains in a β-strand alternately point to either side of the sheet they are consequently alternately buried and exposed to solvent. This structure is reflected in the degree to which they are conserved in Fig. 7 where alternately conserved and mutable positions can be seen in the β-strand regions. The effect is most clearly seen in strands which do not lie on the edge of the β-sheet. In these the conserved positions are generally hydrophobic.

Plotting the degree of conservation with increasing numbers of aligned sequences revealed the interesting observation that many positions rapidly acquire a degree of conservation that is unchanged by the addition of further sequences to the alignment. Conservation is rapidly lost on aligning the first 20 sequences in order of decreasing homology to REI (see Fig. 6(b)) but the addition of another 50 sequences to the alignment causes little alteration of the conservation profile.

FIG. 5. Intersection and union of the sets defined in Fig. 6 can produce a vast number of amino acid combinations. Those which, on an intuitive basis, seem most relevant to protein structure are defined below. These are generally unions and intersections of adjacent sets and thus produce groups of amino acids which share similar properties. Many of the sets have effectively two entries, one with proline and one without. The nomenclature used was chosen to be more readable than standard set notation and uses . or. as the inclusive "or" to represent set union. Intersection of two sets is represented by the underline character "_". Negation is indicated by the prefix "non-" and applies only to the set to which it is attached. This produces recognisable phrases such as "non-POLAR" and "SMALL_HYDROPHOBIC". Occasionally a commonly used set is given a "trivial" name: for example, the set of CHARGED_non-POSITIVE is referred to as "Negative" and the set of HYDROPHOBIC_non-POLAR as "Very-hydrophobic". Some sets which have rather long formal names are simplified by reference to individual amino acids using the one letter code, for example, the set of all hydrophobic residues excluding polar-aromatics is simply called "Very-hydrophobic. or. T. or. K". The ambiguous residue codes asx (b) and glx (z) are included when the two residues they represent occur in the set. They do not, however, count when the number of members in the set is considered.

5. Conclusions

In the rapidly developing field of protein engineering it is important to have a measure of amino acid conservation that can be applied to several homologous sequences of which at least one has a known tertiary structure. General measures of amino acid conservation, such as Dayhoff's likelihood matrix are best suited to situations where there is no structural information about the sequences. Their use becomes limiting when applied to local regions of the protein sequence where, for structural reasons, the mutational freedom of a particular residue may be greatly restrained. With knowledge of the local structure, however, it is possible to analyse these restrictions and use them predictively. An example of loss of information by averaging, which is apparent in Dayhoff's matrix, is the resistance of cyst(e)ine to mutation. This, undoubtedly, arises from the evolutionary need to conserve disulphide bonds, but such a restraint does not apply in the reducing intracellular environment and to apply it to the comparison of the sequences of cytoplasmic proteins is, therefore, misleading.

The classification of amino acids defined above, and its use in describing sequence alignments, allows the type of conservation observed in structural "micro-environments" to be rigorously quantified. The important aspect of the approach is that not only can the degree of conservation be measured, but the qualitative aspect of the conservation is also measured. Together these measures capture virtually all the useful information that can be extracted from a number of aligned sequences. Such information will be of use in designing new mutants as a protein engineer can analyse a sequence alignment to find, for every position, the range of possible amino acid changes that might be acceptable. On a wider front, the approach is being applied to the analysis of residue conservation in well defined structural motifs

FIG. 6. (a) Alignment of the variable domains of the immunoglobulin kappa-chains (light-variable domain) found in the PIR database. The sequences run down the page and are represented in the one letter amino acid code (insertions are indicated by a bar "|"). The numbers identify the position in the sequence of REI and next to these, heavy bars indicate regions of β-structure. To the right of the sequences the amino acid set (see Fig. 5) which best describes the alignment is indicated both by name and number. This description is indented in proportion to the number of insertions. The number just to the left of the set description is the number of amino-acids which constitute the set. This gives a measure of specificity and is the number plotted in Fig. 7. Absolutely conserved positions are indicated by the residue name only. The set assignments are divided into two regions. The outer assignments derive from the 23 sequences with greatest homology to REI while the inner is derived from the entire 73 sequences. The sequence names and their relative homology can be found in (b) which is a matrix of homology between every pair of sequences. The homology is calculated as a percentage of residue identity matches over all matched pairs in two given sequences and is entered in the matrix at a position cross-referenced by the two sequence names. For clarity, homologies over 70% are filled solid, those in the 60s are a dot, those in the 50s a "5" and those below 50% are blanked out. The entries in the matrix have been ordered such that most similar sequences tend to be adjacent on the diagonal. This is achieved by minimising the second moment of homology about the diagonal; i.e.

$$\sum_{i=1}^{N}\sum_{j=1}^{N} H_{ij}(i-j)^2 \rightarrow \min$$

where H_{ij} is the percentage homology between proteins at positions i and j and N is the number of proteins. The increments of ten sequences each corresponding to a plot in Fig. 7 are indicated.

(a)

have a
ologous
easures
uited to
heir use
ere, for
greatly
analyse
tion by
eline to
e disul-
cellular
plasmic

equence
environ-
is that
pect of
/ all the
s. Such
cer can
amino
s being
motifs

-variable
the one
on in the
t of the
by name
r just to
measure
d by the
ive from
quences.
omology
matches
ferenced
re a dot,
ordered
nimising

mber of

(a)

SEQUENCES

73

23

SETS to 73

```
          63 11"SMALL_non-P.or.Hydrophylic"
61 12"SMALL_non-P.or.POLAR_non-AROMATIC"
80  5"ALIPHATIC.or.Large_non-POLAR"
73  8"ALIPHATIC.or.SMALL_non-Hydrophylic"
33  5"Hydrophylic_non-POSITIVE"
88 17"HYDROPHOBIC.or.SMALL"

60 16"SMALL.or.POLAR"

52  7"TINY.or.SMALL_POLAR.or.P"
87 16"Very-Hydrophobic.or.SMALL.or.R"
50  3"TINY"
55 10"TINY.or.POLAR_non-AROMATIC"

70  6"SMALL_non-POLAR_non-P"
86 13"Very-Hydrophobic.or.SMALL_non-P.or.R"
80  5"ALIPHATIC.or.Large_non-POLAR"

CYS
 42  9"POLAR_non-AROMATIC.or.CHARGED"
 51  6"TINY.or.SMALL_POLAR"
 51  6"TINY.or.SMALL_POLAR"
 55 10"TINY.or.POLAR_non-AROMATIC"
 86 13"Very-Hydrophobic.or.SMALL_non-P.or.R"
```

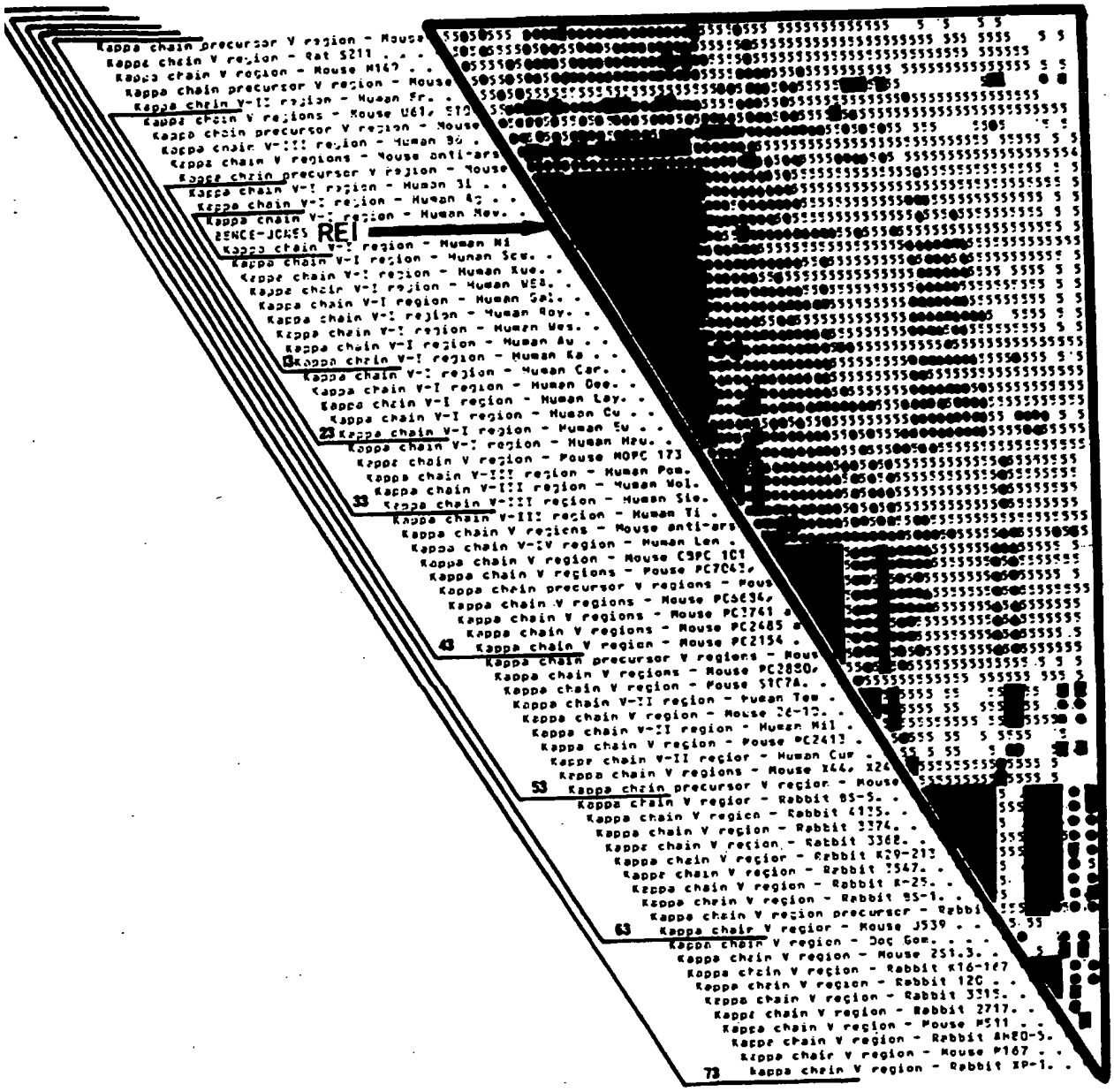
```
 80  5"ALIPHATIC.or.Large_non-POLAR"

TRP
 93  8"AROMATIC.or.ALIPHATIC.or.R"

 97  7"Large_POLAR"
 56 11"TINY.or.POLAR_non-AROMATIC.or.P"
 56  9"TINY.or.Negative_Hydrophylic.or.T.or.P"
 55 10"TINY.or.POLAR_non-AROMATIC"

 56 11"TINY.or.POLAR_non-AROMATIC.or.P"
 87 16"Very-Hydrophobic.or.SMALL.or.R"
 55 10"TINY.or.POLAR_non-AROMATIC"

 93  8"AROMATIC.or.ALIPHATIC.or.R"
 91 16"HYDROPHOBIC.or.P"

 89 16"HYDROPHOBIC.or.SMALL_non-P"

      55 10"TINY.or.POLAR_non-AROMATIC"

 53  8"TINY.or.Negative_Hydrophylic.or.T"
 88 13"Very-Hydrophobic.or.SMALL_non-P.or.R"
 65  5"SMALL_POLAR.or.P"
 55 10"TINY.or.POLAR_non-AROMATIC"
ARG
PHE

  5  6"BEND"
 53  8"TINY.or.Negative_Hydrophylic.or.T"
  5  6"BEND"
 89 16"HYDROPHOBIC.or.SMALL_non-P"
 55 10"TINY.or.POLAR_non-AROMATIC"
 53  8"TINY.or.Negative_Hydrophylic.or.T"
 35  8"CHARGED.or.Hydrophylic"
 27  2"PHEETYR"
 51  6"TINY.or.SMALL_POLAR"
 80  5"ALIPHATIC.or.Large_non-POLAR"

 78  3"ALIPHATIC"
 57 13"TINY.or.POLAR"
 56 11"TINY.or.POLAR_non-AROMATIC.or.P"
 80  5"ALIPHATIC.or.Large_non-POLAR"
 99 11"Large"

 53  8"TINY.or.Negative_Hydrophylic.or.T"
 47  3"SMALL_Hydrophylic"

 86 13"Very-Hydrophobic.or.SMALL_non-P.or.R"
 68  3"SMALL_HYDROPHOBIC"
 89 16"HYDROPHOBIC.or.SMALL_non-P"
TYR
 95  4"AROMATIC"
CYS

 58 19"SMALL_non-P.or.POLAR"

 88 17"HYDROPHOBIC.or.SMALL"
PHE
GLY
 62 13"SMALL.or.POLAR_non-AROMATIC"
 50  3"TINY"
      26  2"SERETHR"

          76  5"ALIPHATIC"
                63 11"SMALL_non-P.or.Hydrophylic"
                86 13"Very-Hydrophobic.or.SMALL_non
                42  9"POLAR_non-AROMATIC.or.CHARGED
```

SETS to 23

```
 33  5"Hydrophylic_non-POSITIVE"
 76  3"ALIPHATIC"
 63 11"SMALL_non-P.or.Hydrophylic"
 80  5"ALIPHATIC.or.Large_non-POLAR"
 73  8"ALIPHATIC.or.SMALL_non-Hydrophylic"
 33  5"Hydrophylic_non-POSITIVE"
 26  2"SERETHR"
 65  5"SMALL_POLAR.or.P"
 86 13"Very-Hydrophobic.or.SMALL_non-P.or.R"
 87 16"Very-Hydrophobic.or.SMALL.or.R"
 80  5"ALIPHATIC.or.Large_non-POLAR"
 65  5"SMALL_POLAR.or.P"
 81  8"Very-Hydrophobic"
 26  2"SERETHR"
 76  7"ALIPHATIC.or.SMALL_non-POLAR"
GLY
 33  5"Hydrophylic_non-POSITIVE"

 70  6"SMALL_non-POLAR_non-P"
 86 13"Very-Hydrophobic.or.SMALL_non-P.or.R"
 80  5"ALIPHATIC.or.Large_non-POLAR"

CYS
 35  8"CHARGED.or.Hydrophylic"
 50  3"TINY"
 50  3"TINY"
 33  5"Hydrophylic_non-POSITIVE"
 51  6"TINY.or.SMALL_POLAR"
 86 13"Very-Hydrophobic.or.SMALL_non-P.or.R"
```

```
 89 16"HYDROPHOBIC.or.SMALL_non-P"
 89 16"HYDROPHOBIC.or.SMALL_non-P"
 76  3"ALIPHATIC"
 86 13"Very-Hydrophobic.or.SMALL_non-P.or.R"
TRP
 93  8"AROMATIC.or.ALIPHATIC.or.R"

 33  5"Hydrophylic_non-POSITIVE"
 55 10"TINY.or.POLAR_non-AROMATIC"
 65  5"SMALL_POLAR.or.P"
  5  6"BEND"

 51  6"TINY.or.SMALL_POLAR"
 76  7"ALIPHATIC.or.SMALL_non-POLAR"
 35  8"CHARGED.or.Hydrophylic"

LEU
 80  5"ALIPHATIC.or.Large_non-POLAR"
 27  2"PHEETYR"

 67  6"SMALL_HYDROPHOBIC.or.TINY"
 64  6"POLAR_non-AROMATIC_non-POSITIVE"

 99 11"Large"
 57 13"TINY.or.POLAR"
 86 13"Very-Hydrophobic.or.SMALL_non-P.or.R"
GLY
 76  3"ALIPHATIC"
PRO
 35  8"CHARGED.or.Hydrophylic"
ARG
PHE

  5  6"BEND"
 51  6"TINY.or.SMALL_POLAR"
  5  6"BEND"
 89 16"HYDROPHOBIC.or.SMALL_non-P"
 50  3"TINY"
 59  6"TINY.or.SMALL_POLAR"
 35  8"CHARGED.or.Hydrophylic"
 27  2"PHEETYR"
 51  6"TINY.or.SMALL_POLAR"
 80  5"ALIPHATIC.or.Large_non-POLAR"
 42  9"POLAR_non-AROMATIC.or.CHARGED"
ILE
 46  6"SMALL_POLAR"
 55 10"TINY.or.POLAR_non-AROMATIC"
 78  3"ALIPHATIC"
 99 11"Large"
 56  9"TINY.or.Negative_Hydrophylic.or.T.or.P"
 33  5"Hydrophylic_non-POSITIVE"
 47  3"SMALL_Hydrophylic"

 80  5"ALIPHATIC.or.Large_non-POLAR"
 70  6"SMALL_non-POLAR_non-P"
 46  6"SMALL_non-P"
TYR
 95  4"AROMATIC"
CYS

 33  5"Hydrophylic_non-POSITIVE"

 57 13"TINY.or.POLAR"
 40 11"POLAR"
 89 16"HYDROPHOBIC.or.SMALL_non-P"
       65  5"SMALL_POLAR.or.P"

 86 13"Very-Hydrophobic.or.SMALL_non-P.or.R"
PHE
GLY
 62 13"SMALL.or.POLAR_non-AROMATIC"
 50  3"TINY"
 26  2"SERETHR"
 42  9"POLAR_non-AROMATIC.or.CHARGED"
 76  3"ALIPHATIC"
 33  5"Hydrophylic_non-POSITIVE"
 86 13"Very-Hydrophobic.or.SMALL_non-P.or.R"
 62  9"POLAR_non-AROMATIC.or.CHARGED"
          5  6"BEND"
```


(facing page 21



(facing page 217)

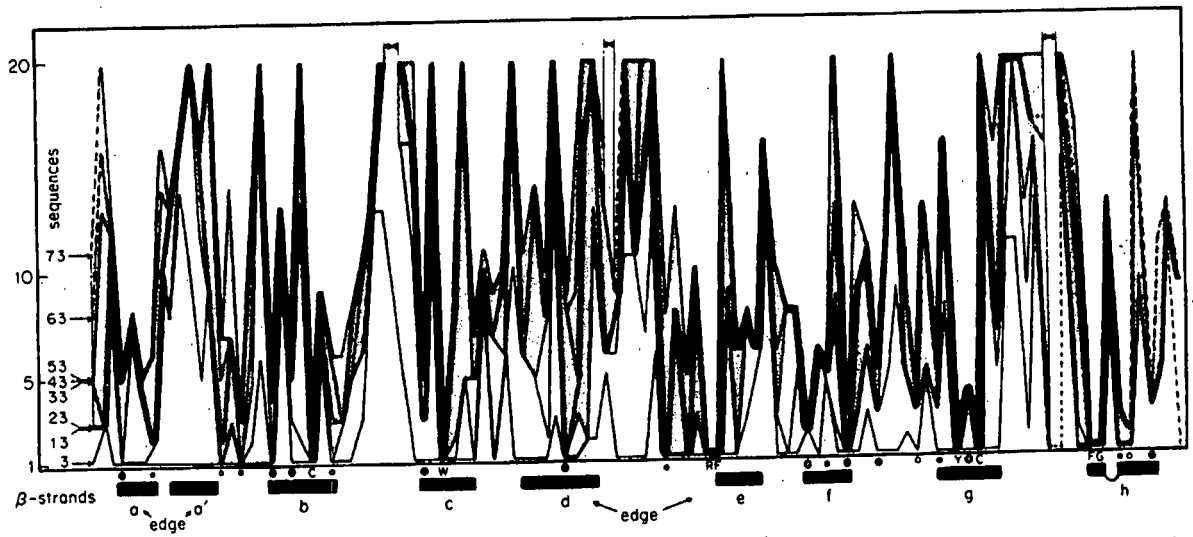


FIG. 7. Plot of degree of conservation along the sequence of aligned immunoglobulin sequences. Conservation is measured by the number of members in the assigned minimal set. This ranges from 1 (absolute conservation of type) to 20 (the set of all amino acids). Eight graphs are plotted, showing the effect of introducing more sequences into the alignment. These progress from three aligned sequences in the steps of ten to the final alignment of 73 sequences. The sequences are added in order of decreasing homology to the Bence Jones protein REI (as measured by residue identity). Insertions required to maintain the alignment occurred in the hyper-variable loops which are indicated by arrows at the top of the plot. The strands of β-structure observed in REI are indicated along the sequence by bars. (Those which lie on the edge of the two β-sheets are also indicated.) Above this, residue types that are conserved in all the sequences are indicated using the one letter code, highly conserved aromatic residues are indicated by a small hexagon, conserved hydrophobic residues by a large black dot, conserved small hydrophobic residues by a smaller dot and conserved small hydrophylic residues by a small open dot (see Fig. 5 for the detailed set assignment). To clarify the graph the line becomes thicker where more than one plot runs together and the region between 73 and 23 sequences is shaded. Considering both these features and bearing in mind that the plots must be monotonic decreasing with the number of sequences, it is possible to identify a single plot at most positions. The dotted lines indicate that a few sequences are not aligned in that region. The sequences are all light-variable domains of the immunoglobulin κ-chains found in the Protein Information Resource databank (Barker *et al.*, 1984). They come from a variety of species but mainly man, rabbit and mouse.

(super-secondary structures) commonly found in globular proteins (Sibanda & Thornton, 1985; Taylor *et al.*, in preparation). In these structures it is found that residue variation is restrained at particular locations in the motifs for general structural reasons. The observed patterns of conservation can then be used to predict the occurrence of the structural motif in a sequence of unknown structure using pattern recognition techniques such as the template matching method of Taylor & Thornton (1983, 1984) and Taylor (1986).

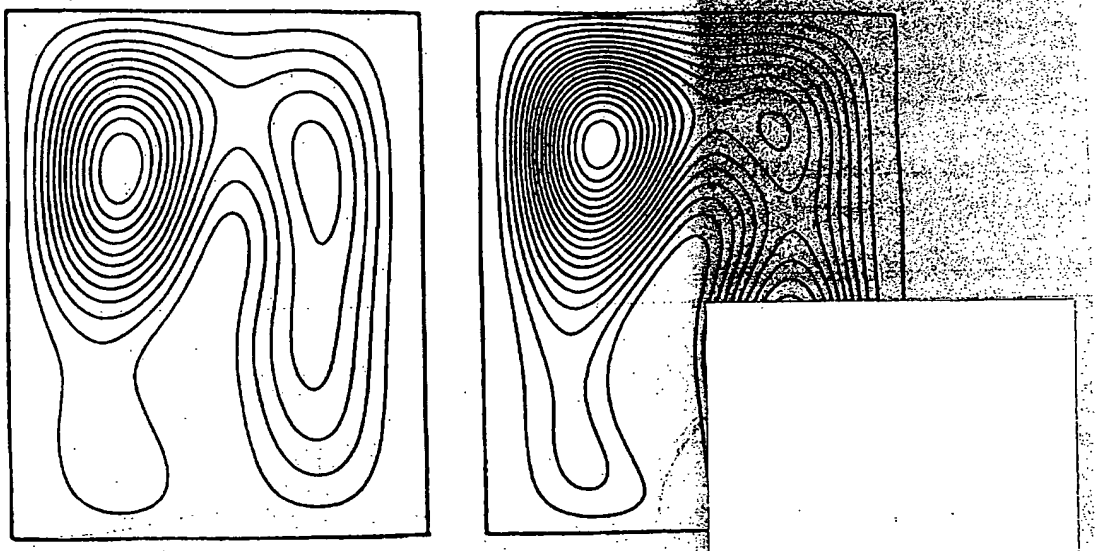
This work was begun while the author was a research fellow at IBM-UK Scientific Centre, Winchester, and the author thanks IBM for support and computing facilities. The author is currently supported by the SERC as an advanced fellow and thanks Prof. T. L. Blundell and Dr J. M. Thornton for valuable discussion.

REFERENCES

BAKER, E. N. & HUBBARD, R. E. (1984). *Prog. Biophys. mol. Biol.* **44**, 97.
BARKER, W. C., HUNT, L. T., ORCUTT, B. C. *et al.* (1984). *Protein Identification Resource.* Release 3.0.
BLUNDELL, T. L., SIBANDA, L. & PEARL, L. (1983). *Nature, Lond.* **304**, 273.
CHOU, P. Y. & FASHMAN, G. D. (1974). *Biochemistry* **13**, 211.
COHEN, F. C., STERNBERG, M. J. E. & TAYLOR, W. R. (1981). *J. mol. Biol.* **148**, 253.
COHEN, F. C., STERNBERG, M. J. E. & TAYLOR, W. R. (1982). *J. mol. Biol.* **156**, 821.
DAYHOFF, M. O. (1972). *Atlas of Protein Sequence and Structure.* Washington DC: National Biomedical Research Foundation.
DAYHOFF, M. O. (1978). *Atlas of Protein Sequence and Structure.* Supplement 3. Washington, DC: National Biomedical Research Foundation.
DICKERSON, R. E. & GEIS, I. (1969). *The Structure and Action of Proteins.* Ch. 1. New York: Harper & Row.
EPP, O., LATTMAN, E. E., SCHIFFER, M., HUBER, R. & PALM, W. (1975). *Biochemistry* **14**, 4943.
FITCH, W. M. (1966). *J. mol. Biol.* **16**, 9.
FRENCH, S. & ROBSON, B. (1983). *J. mol. Evol.* **19**, 171.
GARNIER, J., OSGUTHORP, D. J. & ROBSON, B. (1978). *J. mol. Biol.* **120**, 97.
GRANTHAM, R. (1974). *Science* **185**, 862.
JIMÉNEZ-MONTAÑO, M. A. (1984). *Bull. math. Biol.* **46**, 641.
KLAPPER, M. H. (1971). *Biochim. biophys. Acta* **229**, 557.
MCLACHLAN, A. D. (1971). *J. mol. Biol.* **61**, 409.
MCLACHLAN, A. D. (1972). *J. mol. Biol.* **64**, 417.
REES, A. R. C. & STERNBERG, M. J. E. (1984). *From Cells to Atoms.* Ch. 1. Oxford: Blackwell Scientific.
SANDER, C. & SCHULTZ, G. E. (1979). *J. mol. Evol.* **13**, 245.
SCHULZ, G. E. & SCHIRMER, R. H. (1979). *Principles of Protein Structure.* Ch. 1. New York: Springer-Verlag.
SIBANDA, B. L. & THORNTON, J. M. (1985). *Nature, Lond.* **316**, 107.
SNEATH, D. H. A. (1966). *J. theor. Biol.* **12**, 157.
SWANSON, R. (1984). *Bull. Math. Biol.* **46**, 187.
TAYLOR, W. R. (1981). D. Phil. thesis, University of Oxford.
TAYLOR, W. R. (1986). *J. mol. Biol.* **188** (in press).
TAYLOR, W. R. & THORNTON, J. M. (1983). *Nature, Lond.* **301**, 540.
TAYLOR, W. R. & THORNTON, J. M. (1984). *J. mol. Biol.* **173**, 487.
TAYLOR, W. R., THORNTON, J. M., BARLOW, D. J., SIBANDA, L. & EDWARDS, M. (in preparation).

JOURNAL OF THEORETICAL BIOLOGY

Volume 119 Number 2 21 March 1986



Academic Press
(Harcourt Brace Jovanovich, Publishers)
London Orlando San Diego New York
Toronto Montreal Sydney Tokyo
JTBIAP 119(2) 125–249 (1986) ISSN 0022-5193

J. Mol. Biol. (1991) 217, 721–729

Suggestions for "Safe" Residue Substitutions in Site-directed Mutagenesis

Domenico Bordo[1,2] and Patrick Argos[1]

[1] *European Molecular Biology Laboratory*
Meyerhofstrasse 1, Postfach 10 22 09
6900 Heidelberg, Federal Republic of Germany

[2] *Istituto Nazionale Ricerca sul Cancro*
V. Benedetto XV, 10
16132 Genova, Italy

(Received 19 July 1990; accepted 22 October 1990)

The conserved topological structure observed in various molecular families such as globins or cytochromes *c* allows structural equivalencing of residues in every homologous structure and defines in a coherent way a global alignment in each sequence family. A search was performed for equivalent residue pairs in various topological families that were buried in protein cores or exposed at the protein surface and that had mutated but maintained similar unmutated environments. Amino acid residues with atoms in contact with the mutated residue pairs defined the environment. Matrices of preferred amino acid exchanges were then constructed and preferred or avoided amino acid substitutions deduced. Given the conserved atomic neighborhoods, such natural *in vivo* substitutions are subject to similar constraints as point mutations performed in site-directed mutagenesis experiments. The exchange matrices should provide guidelines for "safe" amino acid substitutions least likely to disturb the protein structure, either locally or in its overall folding pathway, and most likely to allow probing of the structural and functional significance of the substituted site.

1. Introduction

Site-directed mutagenesis has become a very important and yet facile tool to explore the structural and functional significance of particular residues within proteins (for example, see Knowles, 1987; Shaw, 1987; Gruetter *et al.*, 1987). A typical experiment would involve substitutions of an amino acid thought to be essential for catalysis and then assaying the resultant variant for activity. It is central to the success of these experiments that disturbance of the protein fold and structural characteristics, locally as well as globally, be kept to a minimum; otherwise the loss of activity, for instance, would be a result of conformational changes and the exchanged residue be improperly identified as catalytic. Residue substitutions, where the latter situation does not occur, can be considered as "safe".

Natural evolution has "engineered" protein structures by modifying certain molecular properties such as substrate specificity or surface charges and yet conserved the global protein topology. By comparing known conserved three-dimensional protein structures it is possible to glean hints about how this process was performed (Lesk & Chothia, 1980, 1982; Chothia & Lesk, 1986; Bashford *et al.*, 1987); rules obtained in this way are useful for designing site-directed mutagenesis experiments. Protein engineering in the laboratory often faces similar trials. For example, suppose that charges on a protein surface are to be altered to construct a cation binding site. Which amino acids near the surface would be safer to substitute to achieve the desired charge configuration?

In this work residue exchange matrices are calculated that represent point mutational preferences as observed in homologous and known three-dimensional protein structures. Alignments of primary sequences determined from spatial superposition of the main-chain C^{α} and taken from nine molecular families allowed identification of structurally equivalent residues in each of the familial sequence sets. A search was then performed for equivalent residues that had mutated but maintained similar unmutated environments defined by these atoms in contact with the central residue pairs. Such point mutations as observed in known tertiary structures are likely to be, with present-day knowledge, the closest possible mimic of *in vivo* site-directed mutagenesis.

Residue exchange statistics and their significance

0022-2836/91/040721-09 $03.00/0

were determined for all the structural equivalents in the various molecular families. The preferred and avoided substitutions were elicited from three structural contexts: buried residues, amino acids exposed beyond some water-accessible surface area threshold, and then all cases regardless of accessible state. These exchange matrices should provide considerable aid in the difficult process of deciding which residue to exchange and then with which amino acid it should be substituted to maintain protein structural integrity. The preferred exchanges are also discussed in terms of residue physicochemical characteristics.

2. Data and Methods

(a) *Aligned structures*

Aligned sequence sets were taken from 9 molecular families: globins, immunoglobulins, cytochromes *c*, serine proteases, subtilisins, calcium binding proteins, acid proteases, toxins, and virus capsid proteins. The total number of sequences, each with known 3-dimensional structure as contained in the 1989 Brookhaven database collection (Bernstein *et al.*, 1977), was 55. Table 1 lists their database code identification, protein name, species, reference for the 3-dimensional structure, and, where present, reference in which the alignment of the familial sequences used here was determined. The alignments were generally achieved by careful examination of the X-ray crystallographic structures coupled with spatial superposition of the main-chain C^{α} atoms (Rossmann & Argos, 1981). In 3 cases (calcium binding proteins, acid proteases and toxins) structures were superimposed by the present authors using the technique of Rossmann & Argos (Rossmann & Argos, 1976, 1977; Argos & Rossmann, 1979). Due to the increasing number of solved protein structures, many of those used in the present work extracted from the 1989 release of the Brookhaven database were not included in the references showing the familial alignments. These further sequences, indicated by an asterisk in Table 1, were aligned by the authors to the closest family member in both sequence and structure.

When considering statistics for buried residues (solvent-accessible surface area below an upper limit), both constant and variable domains were utilized from the immunoglobulins. However, the variable regions were excluded from the exchange matrix statistics involving surface-exposed amino acids, since large segments of the variable domain loops bind antigens and therefore are subject to special constraints. For a similar reason, side-chains contributing to subunit interface or cofactor contacts were not included in the substitution calculations.

(b) *Similarity of environment*

In a previous paper, Bordo & Argos (1990) carefully defined a measure of similarity (see S'' as given by them in eqns (1) and (3)) between 2 atomic environments surrounding structurally equivalent residues. The same measure is used here. An environment or neighborhood for a residue (called a central residue) is defined by the number of atoms and amino acid types that are within 4·5 Å (1 Å = 0·1 nm) of any side-chain atom in the surrounded residue. The similarity score S is expressed as a fraction and is defined as:

$$S = \frac{\sum_l (\bar{b}_l + \bar{s}_l \delta_l)}{\sum_l (\bar{b}_l + \bar{s}_l)} \qquad (1)$$

The denominator is simply the mean number of atoms belonging to residues present in at least 1 of the 2 environments ($\bar{b}_l$ = main-chain atoms, $\bar{s}_l$ = side-chain atoms). The mean refers to the 2 sets of atoms in each of the 2 environments. The numerator is the sum of the mean number of all main-chain atoms by the 2 environments regardless of the mutational state of the equivalent neighborhood residues plus the mean number of side-chain atoms $\bar{s}_l$ from residues that touch at least 1 atom of the mutated central residues (i.e. within 4·5 Å). The term δ_l is 0 if the lth residue is mutated and 1 if identically conserved. $\sum_l$ is over all residues that touch at least 1 of the central residues. Therefore, similarity of 2 environments will be diminished only if there are mutations in the equivalent environmental residues. That is, if structurally equivalent residues forming the neighborhood of a central residue in 1 protein structure are conserved in the other structure despite their absence in the neighborhood of the equivalent central residue in the latter structure, the similarity score is not decreased. This allows for cases where contacts made by the substituted central residue with its neighbors change only in consequence of its change in size and shape. For instance, environmental residues can move considerably to accommodate a small residue changing to a large one. Though the side-chains in contact with the larger residue are not in contact with the small one, they are nonetheless available without mutation to make contact as necessitated by the substituted residue. Water-accessible surfaces of the combined main-chain and side-chain for each residue was calculated by the procedure of Kabsch & Sander (1983).

(c) *Statistical significance of exchanges*

Counts were made for every observable substitution of central residues with similar neighborhood at a preset similarity threshold. To give statistical significance to these figures, a comparison between observed and expected number of substitutions was performed under the following hypothesis. Consider a pool of N amino acids. $N = \sum_i n_i$ ($i = 1$ to 20), where the ith amino acid type appears n_i times. The exchange $i \rightarrow j$ is a directed replacement of the amino acid i with the amino acid j (e.g. Ala → Asp) and substitution i–j refers to either $i \rightarrow j$ or $j \rightarrow i$ (e.g. Ala → Asp or Asp → Ala). There are $N(N-1)$ possible exchanges in the pool, of which $\sum_i n_i(n_i-1)$ are between residues of the same kind. Therefore, $N' = N(N-1) - \sum_i n_i(n_i-1)$ is the number of possible exchanges involving pairs of different residues. Since the observed mutations refer to only substituted residues, N', and not N, represents the pool of available exchanges. The probability p_{i-j} is then given by $n_i n_j / N'$, and the probability to observe a substitution p_{i-j} becomes:

$$p_{i-j} = 2n_i n_j / N'. \qquad (2)$$

Given a total number of X observed substitutions, the expected number of substitutions n_{i-j} is therefore Xp_{i-j}.

The population n_i ($i = 1$ to 20) was calculated in the following manner. Given a set of structurally aligned sequences for a particular molecular family, each alignment column would generally contain several amino acid types. The count for the population n_i ($i = 1$ to 20) was

Table 1
Tertiary structures used in this work

Family	BRK†	Protein	Origin	Structure reference	Alignment reference‡
Hemoglobin					Lesk & Chothia (1980)
	4HHB	Hemoglobin	Human	Fermi *et al.* (1984)	
	2MHB	Hemoglobin	Equine	Ladner *et al.* (1977)	
	1FDH	Gamma globin	Human	Frier & Perutz (1977)	*
	1MBD	Myoglobin	Whale	Phillips (1980)	
	1MBS	Myoglobin	Seal	Scouloudi & Backer (1978)	*
	2LHB	Hemoglobin V	Sea lamprey	Hendrickson *et al.* (1973)	
	1ECA	Erythrocruorin	*Chironomous*	Steigemann & Weber (1979)	
	2LH1	Leghemoglobin	Lupin	Vainshtein *et al.* (1977)	
Immunoglobulins					Amzel & Poljak (1979)
	1FB4	FAB Kol	Human	Marquart *et al.* (1980)	*
	1FBJ	FAB IgA	Mouse	Navia *et al.* (1979)	*
	1FC1	FcIggl	Human	Deisenhofer (1981)	*
	1FC2	Fc	Human	Deisenhofer (1981)	*
	1IG2	Fc Kol	Human	Marquart *et al.* (1980)	*
	1MCP	FAB	Mouse	Segal *et al.* (1974)	*
	1PFC	Fc Iggl	Porcine	Bryant *et al.* (1985)	*
	1REI	FAB Bence-Jones	Human	Epp *et al.* (1975)	*
	2RHE	FAB Bence-Jones	Human	Furey *et al.* (1983)	*
	3FAB	FAB New	Human	Saul *et al.* (1978)	
	2HFL	FAB Iggl	Mouse	Sheriff *et al.* (1987)	*
	1F19	FAB	Mouse	Lascombe *et al.* (1989)	*
Cytochromes *c*					Dickerson (1980)
	155C	Cytochrome c550	*Paracoccus* D	Timkovich & Dickerson (1976)	
	3C2C	Cytochrome c2	*Rhodospirillum* R	Salemme *et al.* (1973)	
	4CYT	Cytochrome *c*	Bonito fish	Takano & Dickerson (1980)	
	1CYC	Ferrocytochrome *c*	Tuna fish	Tanaka *et al.* (1975)	*
	1CCR	Cytochrome *c*	Rice	Ochi *et al.* (1983)	*
	451C	Cytochrome c551	*Pseudomonas* A	Matsuura *et al.* (1982)	
Serine proteases					Craik *et al.* (1983)
	2SGA	Proteinase A	*Streptomyces* G	Moult *et al.* (1985)	
	3SGB	Proteinase B	*Streptomyces* G	Read *et al.* (1983)	
	2ALP	Alpha-lytic protease	*Lysobacter* E.	Fujinaga *et al.* (1985)	
	4CHA	Alpha chymotrypsin	Bovine	Tsukada & Blow (1985)	
	3PTB	Beta trypsin	Bovine	Marquart *et al.* (1983)	
	2TRM	Trypsin	Rat	Sprang *et al.* (1987)	*
	1TON	Tonin	Rat	Fujinaga & James (1987)	*
	2KAI	Kallikrein	Porcine	Bode *et al.* (1983)	*
	1SGT	Trypsin	*Streptomyces* G	Read & James (1988)	*
	3EST	Elastase	Porcine	Meyer *et al.* (1988)	*
	3RP2	Mast cell protease	Rat	Remington *et al.* (1988)	*
Subtilisins					Froemmel & Sander (1989)
	1SBT	Subtilisin	*B. amylolique-facensis*	Alden *et al.* (1971)	
	2PRK	Proteinase K	Fungus	Paehler *et al.* (1984)	
	1CSE	Subtilisin Karlsberg	*B. subtilis*	Bode *et al.* (1987)	
Calcium binding proteins					
	3CLN	Calmodulin	Rat	Babu *et al.* (1988)	*
	3CPV	Ca-binding parvalbumin B	Carp	Moews & Kretsinger (1975)	*
	3ICB	Ca binding protein	Bovine	Szebenyi & Moffat (1986)	*
	4TNC	Troponin C	Chicken	Satyshur *et al.* (1988)	*
Acid proteases					
	2APP	Penicillopepsin	Fungus	James & Sielecki (1983)	*
	2APR	Rhizopuspepsin	Mold	Suguna *et al.* (1987)	*
	4APE	Endothiapepsin	Fungus	Pearl & Blundell (1984)	*
Toxins					
	1CTX	Alpha cobratoxin	Cobra	Walkinshaw *et al.* (1980)	*
	1NXB	Neurotoxin B	Sea snake	Tsernoglou *et al.* (1978)	*
	2ABX	Alpha bugarotoxin	Krait	Love & Stroud (1986)	*
Viruses					
	2TBV	Tomato bushy stunt	Virus	Hopper *et al.* (1984)	Rossmann *et al.* (1983)
	4SBV	Southern bean mosaic	Virus	Silva & Rossmann (1985)	
	2STV	Satellite tobacco necr.	Virus	Jones & Liljas (1984)	
	1MEV	Mengo	Virus	Luo *et al.* (1987)	Luo *et al.* (1987)
	4RHV	Rhino	Virus	Arnold & Rossmann (1988)	Luo *et al.* (1987)

† The column labeled BRK gives the Brookhaven database entry name (Bernstein *et al.*, 1977).

‡ References showing structural sequence alignments used in this work. An asterisk refers to the cases where the structural alignment was performed by the authors.

Table 2
Residue counts for the nine structural protein families

Residue type	Buried†	Exposed‡	All§
Gly	161	226	445
Ala	182	250	515
Ser	108	375	533
Pro	34	194	249
Asp	28	255	315
Cys	38	23	71
Asn	33	258	313
Thr	79	341	477
Glu	11	239	255
Val	206	166	415
Gln	26	201	248
His	20	69	105
Met	49	47	107
Leu	165	135	331
Ile	125	104	265
Lys	5	297	320
Arg	9	162	193
Phe	89	88	208
Tyr	38	128	191
Trp	30	33	68

† Residues having solvent-accessible surface less than or equal to 10 Å^2. Counts are performed as described in Data and Methods.

‡ Residues having solvent-accessible surface more than or equal to 30 Å^2. Counts are performed as described in Data and Methods.

§ All residues are counted, regardless of their exposure to solvent.

increased by 1 only once for each amino acid type in the alignment column, regardless of its number of appearances. This was consistent with the counts for redundant central residue pairs. For instance, suppose an alignment position contained 3 Ala and 2 Gly residues in a particular topologic family, a total of 6 residue substitutions can be counted; however, since they are all structurally equivalent, only 1 should be taken; namely, that Gly–Ala substitution with the highest environmental similarity score. This selection is consistent with the aim of this study to find conserved neighborhoods tolerating mutant central residues. Total counts n_i ($i = 1$ to 20) were determined for all the alignment positions in all the molecular families under 3 water-accessible conditions and are given in Table 2. The probability to observe α substitutions i–j out of X trials taken from a pool of N residues ($N = \sum_i n_i$) assuming a binomial distribution is given by:

$$P_{i-j}(X,\alpha) = \binom{X}{\alpha} p_{i-j}^{\alpha}(1-p_{i-j})^{X-\alpha}, \qquad (3)$$

where p_{i-j} is given in eqn (2), and:

$$\binom{X}{\alpha} = \frac{X!}{\alpha!\,(X-\alpha)!}.$$

Given the number of observed substitutions n_{i-j}, it is straightforward to calculate its chance probability with eqn (3) (see e.g. Korn & Korn, 1968). If the sum of all probabilities $p_{i-j}(X,\alpha)$ for $n_{i-j} \leq \alpha \leq X$ is less than or equal to 0·05, the preference of the substitutions can be considered significant at the 95% confidence level or better. Consider the following hypothetical illustration. Suppose the pool of residues consisted of 10 amino acids for each of the 20 types ($n_i = 10$, $i = 1$ to 20), then $N = 200$ and the number of possible non-identical amino acid exchanges N' is:

$$(200 \times 199) - \sum_i (10 \times 9) = 38{,}000.$$

If, for instance, 1000 substitutions are observed ($X = 1000$), the expected n_{i-j} using eqn (2), is $2 \times 1000 \times 10 \times 10/38{,}000 \sim 6$. Assume that for a given pair i–j (e.g. Ala–Thr) the observed number of substitutions $n_{\text{Ala-Thr}}$ is 12, then if

$$P_{\text{Ala-Thr}}(1000,12) + P_{\text{Ala-Thr}}(1000,13) + \ldots + P_{\text{Ala-Thr}}(1000,1000) <= 0{\cdot}05$$

the substitution preference between Ala and Thr can be considered significant with at least 95% confidence.

Table 3
Number of substitutions for buried residues involving volume and polarity alterations

Similarity (%)†	100	95	90	85	80
Observed substitutions	12	34	65	124	206
Total number with volume change >1 methyl group	—	1	9	24	57
Total number with polarity group change	2	2	14	33	63
Hydrophobic/hydrophilic substitutions	—	—	1	1	1

† Percentage similarity threshold of central residue environments (see eqn (1)).

3. Results and Discussion

Table 2 lists the residue population for each of the amino acids in the three structural states examined for central residue substitutions: (1) buried in the protein core (solvent-accessible surface for both residues ≤ 10 Å^2); (2) exposed (solvent-accessible surface area ≥ 30 Å^2); and (3) all the possible accessibility states allowed. The residue pool represents, under the constraints discussed in Data and Methods, the composition of amino acids available for possible substitutions. These populations are important in calculating the substitution statistical significance (see Data and Methods).

In a previous paper (Bordo & Argos, 1990), substitution statistics were gathered from only one sequence family (globins) and for only buried residues. The buried exchange counts given here increased by at least a factor of 5 from the addition of eight sequence families (Table 1). The basic trends observed were nonetheless conserved. The results in Table 3 make this salient. Very few of the total substitutions show volume changes greater than one methyl group (~ 35 Å^3) and a movement (referred to as a "jump") to another polarity group (Grantham, 1974) where the three possible groups are defined (1 letter code used) by (WYFMCILV), (PATGS) and (HKRQDEN). These constraints imply considerable impact on the development of protein cores in structures maintaining main-chain fold; a detailed discussion can be found in the earlier work (Bordo & Argos, 1990). All ensuing work given here is unique to this report.

Table 4
Number of substitutions for exposed residues involving volume and polarity alterations

Similarity (%)†	100	95	90	85	80
Observed substitutions	100	152	322	560	941
Total number with volume change >1 methyl group	28	54	124	268	466
Total number with polarity group change	42	69	153	280	547
Hydrophobic/hydrophilic substitutions	3	5	19	39	78

† Percentage similarity threshold of central residue environments (see eqn (1)).

Table 4 lists similar statistics (volume and polarity group alteration counts) for exposed residues with similar environments. It is clear that they display considerable point mutation freedom compared to the buried residues. Approximately one-third to one-half of the substitutions (depending on the percentage similarity of the neighborhood) involve changes in polarity group or volume alterations greater than one methyl group, whereas only about 15% of the buried substitutions involved such changes. However, few side-chains (~5%) alter the sign of their charge or jump (~3%) between opposite polarity groups (i.e. hydrophobic–hydrophilic) despite their exposure.

It was insisted that each of the two substituted residues have a water-accessible surface area of at least 30 Å^2 to be deemed exposed. This represents approximately a hole just large enough for a methyl group to pass through and was found from the previous globin statistics (Bordo & Argos, 1990) as well as the present data (not shown) to be the minimal exposure at which radical volume and polar alterations between exchanged central residues are observed.

Figure 1 shows the actual exchange counts for (a) buried, (b) exposed and (c) all cases where the central residue environments were 90% or greater (lower matrix half) and 70% or greater (upper matrix half) in similarity. The symbols plus (preferred exchange) and minus (avoided exchange) are shown in the upper half of matrices if the counts were reliable at the 95% confidence level or better as well as consistently preferred or shunned for at least two similarity levels within a range of 100% to 70% calculated in steps of 5%. As expected, the 70% similarity data produced the most observed exchange counts and the greatest number of substitutions deemed significant. However, given the lessened neighborhood similarity, noise is increasingly introduced; nonetheless, trends are preserved from the 90% to 70% levels (Fig. 1).

Several interesting substitution trends are observable in the Figure 1 exchange matrices. Though the high count substitutions are not always deemed statistically significant, they represent a useful starting point in deciding which substitutions to try in structure-altering experiments as site-directed mutagenesis or protein engineering. It will take considerable time and effort to produce sufficient X-ray crystallographic protein structures to determine the significance of all the possible substitutions.

For the protein core, residues within each of the following subsets are generally interchangeable with high statistical significance: (A, G), (A, V), (N, D), (M, L), (F, L), (F, Y), (A, S, T), (V, I, L) and (Y, W). This is shown diagrammatically in Figure 2. In an examination of the counts alone, surprising results can be found for many of the amino acid types. While Thr can exchange with Ala and Ser, Asn is the next most desirable. Cys prefers Ala or Val as substitutes. Though Val can rather freely go to Ala, Ile and Leu, Ile prefers primarily only Val and Leu. Met and Phe favor Leu, rather than Ile, as an ersatz. For exposed substitutions unexpected results are also in evidence. Gly prefers Asn as the most desirable charged or polar substitute. If Ala must be replaced by a charged residue, Lys and Glu are statistically favored. Ser prefers Asp and Asn and not Glu, Lys or Arg, while Thr is the most favored substitute. Asp especially avoids Tyr at the surface. Val's favorite partners are Ile and Leu, while Tyr prefers Phe. Interestingly, the hydrophobic residues Val, Leu and Ile tend to substitute amongst themselves despite some exposure at the surface. If an exposed Val must be changed to a charged residue, Lys is the best candidate; and so forth.

Some substitutions are consistently allowed regardless of exposure or buriedness (Fig. 2). Among the highly significant preferred exchanges, in single letter code, are (G, A), (S, A), (T, A), (N, D), (T, S), (V, I, L) and (F, Y).

Calculating the logarithm of the ratio of the observed to expected counts for each possible substitution and for all observed cases having 70% environmental similarity (Fig. 1(c), upper right matrix), it was possible to build a scoring matrix analogous to that determined by Dayhoff *et al.* (1978). The correlation coefficient between the elements of the two matrices was 0·64. It would not be expected that the two matrices correlate well as the results of this work concern single substitutions over only close molecular generations, while the Dayhoff *et al.* observations are cumulative over many and multiple mutations.

The matrices listing preferred or safe and avoided or unsafe substitutions taken from actual tertiary structures should prove exceedingly useful in site-directed mutagenesis and protein engineering experiments. It would be helpful to ascertain if a residue is exposed or buried before choosing a substitution. If the protein three-dimensional structure is known, this information is evident. If only the sequence has been determined, secondary structure prediction and/or a hydrophobicity plot (for a review, see Argos, 1990) should provide a good guess as to the appropriate solvent-accessible state of the residue in question. If not, the exchange counts taken from all residues in the familial sequence sets are given in Figure 1(c).

	G	A	S	C	T	P	D	V	N	L	I	Q	M	E	H	K	F	R	Y	W
G		23+	6	3	5	0	0	8−	0	4	3	1	1	0	0	0	1−	0	0	0
A	9		18+	4	13+	3	1	32+	3	14	8	3	2	1	0	0	5	1	2	0
S	1	5		1	16+	1	2	5	8	4	3	0	1	0	0	0	3	0	1	1
C	0	1	0		0	0	0	4	0	0	1	0	0	0	0	0	1	0	0	0
T	0	4	5	0		1	1	10	2	3	6	0	4	1	0	0	2	0	0	0
P	0	0	0	0	0		0	1	0	0	0	0	0	0	0	0	1	0	0	0
D	0	0	0	0	0	0		0	3+	0	0	0	0	0	0	0	0	0	0	0
V	1	2	1	1	1	0	0		0	34+	39+	0	11	0	3	0	6	0	3	0
N	0	0	1	0	1	0	1	0		1	0	2	1	0	0	0	1	0	1	0
L	0	1	0	0	0	0	0	3	0		19+	2	15+	0	1	0	13+	0	4	5
I	0	1	0	0	1	0	0	10	0	3		2	4	0	0	0	4	1	1	4
Q	0	0	0	0	0	0	0	0	0	0	0		1	1	0	0	1	0	1	1
M	0	0	0	0	1	0	0	1	1	2	1	0		1	1	0	3	0	1	1
E	0	0	0	0	0	0	0	0	0	0	0	1	0		0	0	0	0	0	0
H	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	1	1	0	0
K	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	1	0	0
F	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0		0	6+	1
R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0		1	0
Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0		4+
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	

(a)

	G	A	S	C	T	P	D	V	N	L	I	Q	M	E	H	K	F	R	Y	W
G		29+	33+	1	20	19	23	3	30+	1−	1−	7	1	14	4	23	1	7	5	1
A	8		53+	2	30+	19+	23	6	26	5	5	20	2	32+	7	34+	3	9	4−	2
S	10	17		1	80+	29	45	13	50	6−	4	34	2	25	14	35	2−	20	4−	0
C	0	0	0		2	1	0	0	0	2	1	0	0	1	0	0	0	0	0	0
T	2	9	21	0		20	20	17	34	14	10	30	5	25	7	38+	7	17	7−	2
P	3	5	6	0	3		21	6−	12	3−	2	11	1	19	8	21	0	6	1	1
D	2	2	9	0	1	5		7−	45+	6	4	17	1	42+	4	25−	2	5	2−	0
V	1	2	2	0	3	0	0		6	14+	14+	7	1	10	0	15	2	8	4	0
N	7	3	7	0	3	0	11	1		10	5	18	1	13	6	32	5	7	6−	2
L	0	1	1	0	1	0	1	2	2		9+	13	4	8	1	14	3	7	7	3
I	0	1	1	0	2	0	0	2	0	4		6	3	7	1	7	6	2	3	1
Q	0	3	4	0	6	1	4	1	3	1	0		5	23+	6	29+	1	11	3	0
M	0	0	1	0	0	0	0	0	0	1	1	1		1	1	3	2	1	0	0
E	3	6	5	0	4	4	10	1	2	1	1	3	0		3	30	3	6	5	0
H	2	1	0	0	1	1	0	0	1	0	0	2	0	0		9	2	1	4	1
K	4	10	2	0	12	2	2	2	6	3	0	5	0	6	0		4	28+	3−	3
F	0	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0		1	11+	3
R	1	1	3	0	1	0	1	2	1	0	0	1	0	0	1	7	0		3	2
Y	0	1	0	0	0	0	0	1	0	2	1	0	0	0	0	1	3	0		4
W	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	

(b)

	G	A	S	C	T	P	D	V	N	L	I	Q	M	E	H	K	F	R	Y	W
G		79+	65	4	45	25	34	16−	44+	9−	7−	19	6	23	8	31	4−	13	10−	3
A	23		110+	8	73+	35	33	63+	39	30−	24	39	11	46+	12	54	10−	21	14	4
S	14	35		5	128+	40	59+	31−	74+	17−	16	46	6−	32	16	53	8−	31	11−	4
C	0	0	0		7	4	0	6	0	3	3	1	0	2	0	0	2	0	0	1
T	7	21	43	0		31	29	48	47+	32	29	42	15	34	11	58+	15	29	10−	1
P	5	10	7	0	7		25	12−	16	7−	3−	11	1	22	8	22	3	9	4−	1
D	7	3	11	0	4	5		12−	61+	9−	9−	19	1−	47+	4	25	4−	9	2−	0
V	2	10	6	1	8	1	2		9−	65+	77+	13	19	18	7	21	19	16	12	0−
N	9	6	12	0	5	0	15	1		13−	8−	20	3	17	8	36	7−	10	9−	2
L	0	2	1	0	3	1	0	7	2		45+	21	30+	15	4	17	22	11	14	9
I	0	3	2	0	7	0	2	17	0	10		8	9	9	2−	7−	15	6−	6	5
Q	1	6	6	0	8	1	4	4	3	1	0		9	29+	10	38+	2−	18	5	2
M	0	0	1	0	4	0	0	2	1	7	3	1		3	3	6	5	2	2	1
E	4	6	6	0	8	5	12	3	2	1	2	9	0		4	31	3−	10	5−	0
H	3	2	0	0	0	2	0	0	1	0	0	0	0	1		8	5	3	8	2
K	5	11	5	0	14	2	2	2	7	2	0	8	0	6	1		4−	38+	3−	3
F	0	0	0	0	1	0	0	0	0	4	2	0	1	0	0	0		4−	23+	9+
R	2	3	4	0	5	1	2	2	2	1	0	4	0	0	2	9	0		4	2
Y	0	2	0	0	0	0	0	2	0	2	1	1	0	0	1	1	5	0		12+
W	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	1	0	1	

(c)

Figure 1. Observed substitutions for (a) buried, (b) exposed and (c) all cases. The lower halves of the matrices give substitution counts for central residues with 90% or greater similar environments, while the upper halves are for 70% or greater similarity. When counts show a statistically meaningful (95% or greater confidence) increase or decrease compared to the expected figures for at least 2 similarity levels ranging from 100% to 70% in steps of 5%, with the trend being consistent, a + or − sign is given to indicate preferred or avoided substitutions, respectively. In the exposed data, immunoglobulin variable domains were not included.

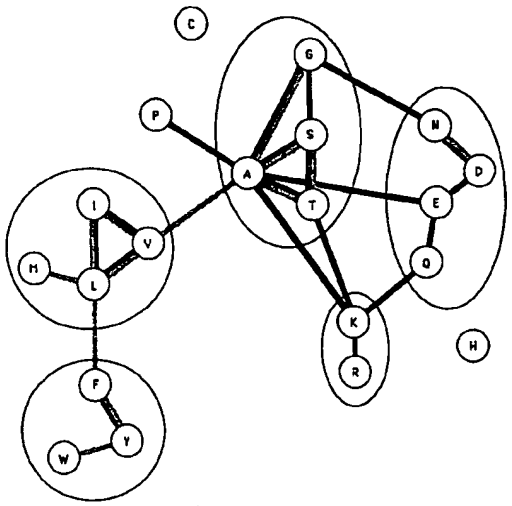


Figure 2. Statistically preferred (95% or greater confidence level as indicated by a + in Fig. 1) substitutions observed in buried residues (grey segments) and exposed residues (black segments) are shown. Residues roughly equivalent are grouped together in 5 subsets, which generally correlate with side-chain physicochemical properties.

Lim & Sauer (1989) have performed mutation experiments on λ repressor protein core side-chains and the mutants were assayed for functionality and stability. Interestingly, all of the single protein core mutants could have been predicted from this work (Bordo & Argos, 1990).

Site-directed mutagenesis is an important tool in probing the structural and functional significance of particular residues within a protein sequence (for reviews, see Knowles, 1987; Shaw, 1987). Amino acid residues might be altered to check for their participation in catalysis, cofactor or substrate binding, molecular and receptor recognition, domain interfaces, oligomeric interactions, and the like. It is essential in such experiments that the protein fold, locally and globally, not be perturbed; otherwise, loss of activity or whatever aspect is under study would be incorrectly ascribed to the mutated residue. "Safe" substitutions are thus requisite for the success of the mutant probe as an indicator of critical residues in structure and function. This work provides exchange matrices that should be directly applicable in maintaining the fold and that are taken from known three-dimensional protein structures with diverse folds. Of course, the results represent general trends and cannot be expected to work in every local context, but they should be a great improvement over randomly selected substitutions and act as a good guide regarding what to substitute and what not to substitute. For example, suppose Cys were a suspected active site residue. If exposed or buried, though the substitution data base is not sufficient to identify statistically significant exchanges for Cys, the observed substitutions counts would recommend Ala; if the Cys is likely to be buried, Val is also a possible candidate.

Zvelebil & Sternberg (1988) examined several known tertiary structures and determined that His is the most frequently occurring catalytic residue. Assuming its exposure to the solvent, the exchange matrix suggests Ser as the safest substitution. In the review by Shaw (1987) on specific point mutations for several molecular species, the Gly–Ala substitution is one of the most frequent mentioned. Apparently the proteins maintained their fold while proven assays displayed altered activity. The exchange matrices presented in this work suggest the Gly–Ala substitution as highly significant in the buried or exposed states.

In protein engineering as well as molecular modeling, where new structures are built from those with known tertiary and homologous primary structures (for a review, see Sali *et al.*, 1990), it is often crucial to know which residues can be substituted safely. Can a substituted residue in a molecular model be placed in the same environment displayed by the known native structure? For instance, if a His is to be introduced in an exposed loop to engineer cation binding, would it be safer to substitute a Ser, Glu, Asn or Lys in the known structure? The exchange matrices of Figure 1 provide direct answers. In fact, Sali *et al.* (1990) in their review on modeling cite only two specific examples where residues are allowed limited choices due to folding requirements. Both involve constrained Ser–Thr substitutions in buried β-strands where the side-chain oxygen atoms bond to main-chain atoms. Among the preferred exchanges, the Ser–Thr one is highly preferred both in the exposed and buried substitutions matrices reported here (Fig. 2). A further protein engineering example would involve a desired residue substitution to stabilize a predicted or known helix. The exchange should be from a residue of lower to higher helical preference (Palau *et al.*, 1982). Combining this requirement with the exchange matrix counts of Figure 1 should provide a very rational substitution, especially if the tertiary structure is not known, which is typically the situation. For example, if Ile were buried and part of a helix is to be stabilized, the matrix of Figure 1(a) suggests Leu and then Met as likely substitution candidates.

Malcolm *et al.* (1990) have published results of mutants of game bird lysozymes. Point mutations on *in vivo* triplets Thr40-Ile55-Ser91 (TIS) or Ser40-Val55-Thr91 (SVT) included, respectively, TVS, SIS, TIT and SVS, SIT, TVT. The mutants were assayed for thermal stability and it was found that TIT, SIT and TVT were more stable than the respective wild-type and TVS, SIS and SVS less so. The buried-residue exchange matrices in this work would predict that Val → Ile and Ser → Thr would be ideal substitutions to preserve main-chain fold and enhance thermal stability under the assumption that increasing the volume of a side-chain within one methyl group would result in better hydrophobic packing to maintain the protein structure. In every case, this is exactly what occurred experimentally. In fact, when the exchange from the wild-

type involved a volume decrease, the fold was maintained but thermal stability diminished.

The authors thank Gareth Chelvanayagam, Jaap Heringa and Peter Sibbald for many helpful discussions.

References

Alden, R. A., Birktoft, J. J., Kraut, J., Robertus, J. D. & Wright, C. S. (1971). *Biochem. Biophys. Res. Commun.* **45**, 337–449.

Amzel, L. M. & Poljak, R. (1979). *Annu. Rev. Biochem.* **48**, 961–997.

Argos, P. (1990). *Methods Enzymol.* **182**, 751–776.

Argos, P. & Rossmann, M. G. (1979). *Biochemistry*, **18**, 4951–4960.

Arnold, E. & Rossmann, M. G. (1988). *Acta Crystallogr. sect. A*, **44**, 270–282.

Babu, Y. S., Bugg, C. E. & Cook, W. J. (1988). *J. Mol. Biol.* **204**, 191–204.

Bashford, D., Chothia, C. & Lesk, A. M. (1987). *J. Mol. Biol.* **196**, 199–216.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Schimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.

Bode, W., Chen, Z., Bartels, K., Kutzbach, C., Schmidt-Kastner, G. & Bartunik, H. (1983). *J. Mol. Biol.* **164**, 237–282.

Bode, W., Papamokos, E., & Musil, D. (1987). *Eur. J. Biochem.* **166**, 673–692.

Bordo, D. & Argos, P. (1990). *J. Mol. Biol.* **211**, 975–988.

Bryant, S. H., Amzel, L. M., Phizackerley, R. P. & Poljak, R. J. (1985). *Acta Crystallogr. sect. B*, **41**, 362–368.

Chothia, C. & Lesk, A. M. (1986). *EMBO J.* **5**, 823–826.

Craik, C. S., Rutter, W. R. & Fletterick, R. (1983). *Science*, **220**, 1125–1129.

Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. (1978). *Atlas of Protein Sequence and Structure*, vol. 5, suppl. 3, pp. 345–362. National Biochemical Foundation, Georgetown University Medical Center, Washington, DC.

Deisenhofer, J. (1981). *Biochemistry*, **20**, 2361–2370.

Dickerson, R. E. (1980). *Sci. Amer.* **242**, 98–112.

Epp, O., Lattman, E. E., Schiffer, M., Huber, R. & Palm, W. (1975). *Biochemistry*, **14**, 4943–4952.

Fermi, G., Perutz, M. F., Shaanan, B. & Fourme, R. (1984). *J. Mol. Biol.* **175**, 159–174.

Frier, J. A. & Perutz, M. F. (1977). *J. Mol. Biol.* **112**, 97–112.

Froemmel, C. & Sander, C. (1989). *Proteins*, **5**, 22–37.

Fujinaga, M. & James, M. N. G. (1987). *J. Mol. Biol.* **195**, 373–396.

Fujinaga, M., Delbaere, L. T. J., Brayer, G. D. & James, M. N. G. (1985). *J. Mol. Biol.* **84**, 479–502.

Furey, W., Jr, Wang, B. C., Yoo, C. S. & Sax, M. (1983). *J. Mol. Biol.* **167**, 661–692.

Grantham, R. (1974). *Science*, **185**, 862–864.

Gruetter, M. G., Gray, T. M., Weawer, L. H., Alber, T., Wilson, K. & Matthews, B. W. (1987). *J. Mol. Biol.* **197**, 315–329.

Hendrickson, W. A., Love, W. E. & Karle, J. (1973). *J. Mol. Biol.* **74**, 331–361.

Hopper, P., Harrison, S. C. & Sauer, R. T. (1984). *J. Mol. Biol.* **177**, 701–713.

James, M. N. G. & Sielecki, A. R. (1983). *J. Mol. Biol.* **163**, 299–361.

Jones, T. A. & Liljas, L. (1984). *J. Mol. Biol.* **177**, 735–767.

Kabsch, W. & Sander, C. (1983). *Biopolymers*, **22**, 2577–2637.

Knowles, J. R. (1987). *Science*, **236**, 1252–1258.

Korn, G. A. & Korn, T. M. (1968). *Mathematical Handbook for Scientists and Engineers*, pp. 10–11, McGraw-Hill Book Company, New York.

Ladner, R. C., Heidner, E. G. & Perutz, M. F. (1977). *J. Mol. Biol.* **114**, 385–414.

Lascombe, M. B., Alzari, P. M., Boulot, G., Saludjian, P., Tougard, P., Berek, C., Haba, S., Rosen, E. M., Nisonoff, A. & Poljak, R. J. (1989). *Proc. Nat. Acad. Sci., U.S.A.* **86**, 607–611.

Lesk, A. M. & Chothia, C. (1980). *J. Mol. Biol.* **136**, 225–270.

Lesk, A. M. & Chothia, C. (1982). *J. Mol. Biol.* **160**, 325–342.

Lim, W. A. & Sauer, R. T. (1989). *Nature (London)*, **339**, 31–36.

Love, R. A. & Stroud, R. M. (1986). *Protein Eng.* **1**, 37–46.

Luo, M., Vriend, G., Kamer, G., Minor, I., Arnold, E., Rossmann, M. G., Boege, U., Scraba, D. G., Duke, G. M. & Palmenberg, A. C. (1987). *Science*, **235**, 182–191.

Malcom, B. A., Wilson, K. P., Matthews, B. W., Kirsh, J. F. & Wilson, A. C. (1990). *Nature (London)*, **345**, 86–89.

Marquart, M., Deisenhofer, J., Huber, R. & Palm, W. (1980). *J. Mol. Biol.* **141**, 369–391.

Marquart, M., Walter, J., Deisenhofer, J., Bode, W. & Huber, R. (1983). *Acta Crystallogr. sect. B*, **39**, 480–490.

Matsuura, Y., Takano, T. & Dickerson, R. E. (1982). *J. Mol. Biol.* **156**, 389–409.

Myer, E., Cole, G., Radhakrishnan, R. & Epp, O. (1988). *Acta Crystallogr. sect. B*, **44**, 26–38.

Moews, P. C. & Kretsinger, R. H. (1975). *J. Mol. Biol.* **91**, 201–228.

Moult, J., Sussman, F. & James, M. N. G. (1985). *J. Mol. Biol.* **182**, 555–566.

Navia, M. A., Segal, D. M., Padlan, E. A., Davies, D. R., Rao, N., Rudikoff, S. & Potter, M. (1979). *Proc. Nat. Acad. Sci., U.S.A.* **76**, 4071–4074.

Ochi, H., Hata, Y., Tanaka, N., Kakudo, M., Sakurai, T., Aihara, S. & Morita, Y. (1983). *J. Mol. Biol.* **166**, 407–418.

Paehler, A., Banerjee, A., Dattagupta, J. K., Fujiwara, T., Lindner, K., Pal, G. P., Suck, D., Weber, G. & Saenger, W. (1984). *EMBO J.* **3**, 1311–1314.

Palau, J., Argos, P. & Puigdomenech, P. (1982). *Int. J. Protein Pept. Res.* **91**, 394–401.

Pearl, L. & Blundell, T. (1984). *FEBS Letters*, **174**, 96–111.

Phillips, S. E. V. (1980). *J. Mol. Biol.* **142**, 531–554.

Read, R. J. & James, M. N. G. (1988). *J. Mol. Biol.* **200**, 523–551.

Read, R. J., Fujinaga, M., Sielecki, A. R. & James, N. M. G. (1983). *Biochemistry*, **22**, 4420–4433.

Remington, S. J., Woodbury, R. G. & Reynolds, R. A. (1988). *Biochemistry*, **27**, 8097–8105.

Rossmann, M. G. & Argos, P. (1976). *J. Mol. Biol.* **105**, 75–95.

Rossmann, M. G. & Argos, P. (1977). *J. Mol. Biol.* **109**, 99–129.

Rossmann, M. G. & Argos, P. (1981). *Annu. Rev. Biochem.* **50**, 497–532.

Rossman, M. G., Abad-Zapatero, C., Murthy, M. R. N.,

Liljas, L., Jones, T. A. & Strandberg, B. (1983). *J. Mol. Biol.* **165**, 711–736.
Salemme, F. R., Freer, S. T., Xuong, N. H., Alden, R. A. & Kraut, J. (1973). *J. Biol. Chem.* **248**, 3910–3921.
Sali, A., Overington, J. P., Johnson, M. S. & Blundell, T. L. (1990). *Trends Biochem. Sci.* **15**, 235–240.
Satyshur, K. A., Sambhrao, S. T., Pyzalska, D., Drendel, W., Greaser, M. & Sundaralingan, M. (1988). *J. Biol. Chem.* **263**, 1628–1647.
Saul, F. A., Amzel, L.M. & Poljak, R. J. (1978). *J. Biol. Chem.* **253**, 585–597.
Scouloudi, H. & Backer, E. N. (1978). *J. Mol. Biol.* **126**, 637–660.
Segal, D. M., Padlan, E. A., Cohen, G. H., Rudikoff, S., Potter, M. & Davies, D. R. (1974). *Proc. Nat. Acad. Sci., U.S.A.* **71**, 4298–4302.
Shaw, W. V. (1987). *Biochem. J.* **246**, 1–17.
Sheriff, S., Silverton, E. W., Padlan, E. A., Cohen, G. H., Smith-Gill, S. J., Finzel, B. C. & Davies, D. R. (1987). *Proc. Nat. Acad. Sci., U.S.A.* **84**, 8075–8079.
Silva, A. M. & Rossmann, M. G. (1985). *Acta Crystallogr. sect. B*, **41**, 147–157.
Sprang, S., Standing, T., Fletterick, R. J., Stroud, R. M., Finer-Moore, J., Xuong, N. H., Hamlin, R., Rutter, W. J. & Craik, C. S. (1987). *Science*, **237**, 905–909.
Steigemann, W. & Weber, E. (1979). *J. Mol. Biol.* **127**, 309–338.
Suguna, K., Bott, R. R., Padlan, E. A., Subramanian, E., Sheriff, S., Cohen, G. H. & Davies, D. R. (1987). *J. Mol. Biol.* **196**, 877–900.
Szebenyi, D. M. & Moffat, K. (1986). *J. Biol. Chem.* **261**, 8761–8777.
Takano, T. & Dickerson, R. E. (1980). *Proc. Nat. Acad. Sci., U.S.A.* **77**, 6371–6375.
Tanaka, N., Yamane, T., Tsukihara, T., Ashida, T. & Kakudo, M. (1975). *J. Biochem.* **77**, 147–162.
Timkovich, R. & Dickerson, R. E. (1976). *J. Biol. Chem.* **251**, 4033–4046.
Tsernoglou, D., Petsko, G. A. & Hudson, R. A. (1978). *Mol. Pharmacol.* **14**, 710–716.
Tsukada, H. & Blow, D. M. (1985). *J. Mol. Biol.* **184**, 703–711.
Vainshtein, B. K., Arutyunyan, E. G., Kuranova, I. P., Borisov, V. V., Sosfenov, N. I., Pavlovskii, A. G., Grebenko, A. I., Konareva, N. V. & Nekrasov, Y. V. (1977). *Dokl. Biochem.* (English translation), **233**, 67–70.
Walkinshaw, M. D., Saenger, W. & Maelicke, A. (1980). *Proc. Nat. Acad. Sci., U.S.A.* **77**, 2400–2404.
Zvelebil, M. J. J. M. & Sternberg, M. J. E. (1988). *Protein Eng.* **2**, 127–138.

Edited by A. Fersht